De nouveaux facteurs pour l'exploitation de la sémantique d'un texte en Recherche d'Information

Ihab Mallak

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par *l'Université Toulouse III - Paul Sabatier*
Discipline ou spécialité : *Informatique et applications*

Présentée et soutenue par *Ihab Mallak*
Le *11 Juillet 2011*

Titre : *De nouveaux facteurs pour l'exploitation de la sémantique d'un texte en Recherche d'Information*

JURY

Claude CHRISMENT Professeur à l'université de Paul Sabatier (Président)
Sylvie CALABRETTO Professeur à LIRIS-INSA, Lyon (Rapporteur)
Gabriella PASI Professeur à l'université di Milano Bicocca, Italie (Rapporteur)
David HAWKING Professeur adjoint à l'université National d'Australie, Australie (Examinateur)
Mohand BOUGHANEM Professeur à l'université de Paul Sabatier
Henri PRADE Directeur de recherche CNRS

Ecole doctorale : *Ecole Doctorale Mathématiques Informatique Télécommunication de Toulouse*
Unité de recherche : *INSTITUT DE RECHERCHE EN INFORMATIQUE DE TOULOUSE -IRIT-*
Directeurs de Thèse : *Boughanem Mohand/ Prade Henri*

Résumé

Les travaux présentés dans ce mémoire se situent dans le contexte de la recherche d'information. Plus précisément, nous proposons de nouveaux facteurs « *centralité, fréquence conceptuelle*» permettant à notre sens, de mieux caractériser la dimension sémantique du contenu des textes, allant au-delà des méthodes d'indexation classiques basées exclusivement sur les statistiques. Ces facteurs devraient tirer parti de l'identification de différents types de relations telles que -est-une partie-de, liés à, synonymie, domaine, etc.- qui existent entre les mots d'un texte.

L'approche que nous avons proposée pour calculer la valeur de nos facteurs est bâtie en trois étapes : (1) Extraction des concepts issus de WordNet[1] associés aux termes du document puis désambigüisation de leurs sens, (2) Regroupement des concepts pour former des clusters de concepts (Ces étapes construisent la vue sémantique des documents), (3) A l'intérieur de chaque cluster, chaque terme possède un degré de « *centralité* », fonction du nombre de mots du cluster avec lequel il est en relation directe, et une « *fréquence conceptuelle* » estimée par la somme des fréquences de ces mots.

D'une part, nous menons une étude sur des méthodes potentielles basées sur les facteurs proposés pour extraire des vues sémantiques du contenu des textes. L'objectif est de construire des structures de graphes/hiérarchies offrant une vue du contenu sémantique des documents. Ensuite, ces vues seront élaborées à partir de nos nouveaux facteurs, mais aussi de l'utilisation des fréquences d'occurrence, et de la prise en compte de l'importance des mots (en particulier en terme de leur spécificité). Le poids relatif des vues partielles, la fréquence et la spécificité de leurs composants sont d'autant des indications qui devraient permettre d'identifier et de construire des sous-ensembles hiérarchisés de mots (présents dans le texte ou sémantiquement associés à des mots du texte), et de refléter les concepts présents dans le contenu du texte. L'obtention d'une meilleure représentation du contenu sémantique des textes aidera à mieux retrouver les textes pertinents pour une requête donnée, et à donner une vue synthétisée du contenu des textes proposés à l'utilisateur en réponse à sa requête.

D'autre part, nous proposons une technique de désambiguïsation du concept basée sur la centralité. En fait, le sens d'un terme est ambigu, il dépend de son contexte d'emploi. Dans notre proposition, nous utilisons l'ontologie de WordNet, qui est précise dans la couverture des sens de termes, où un terme peut être attaché à plusieurs concepts. La méthode proposée consiste à trouver le meilleur concept WordNet permettant de représenter le sens du terme désigné par le texte. Le concept choisi est celui qui a un maximum de relations avec les termes du document, autrement dit, celui qui a une valeur maximale de centralité. L'utilisation d'une méthode de désambiguïsation est une étape inévitable dans une indexation conceptuelle, elle permet de mieux représenter le contenu sémantique d'un document.

Enfin, nous utilisons nos facteurs dans le cadre de Recherche d'Information comme de nouveaux facteurs pour mesurer la pertinence d'un document vis-à-vis d'une requête (tâche de RI ad-hoc). L'utilisation de nos facteurs sémantiques est intéressante dans la RI, où nous estimons un degré de relativité entre les termes d'une requête et ceux d'un document indépendamment de leur présence dans ce dernier. Dans ce cadre, nous avons proposé une nouvelle fonction de pondération basée sur la centralité, ainsi que nous avons intégré les nouveaux facteurs à des fonctions connues.

---

1 Une ressource sémantique contenant un ensemble de concepts et de relations dénotant des liens sémantiques entre ces concepts. http://wordnet.princeton.edu/

Dans les différentes expérimentations menées, nous avons montré que l'intégration de nos facteurs sémantiques ramène une amélioration au niveau de précision dans un moteur de recherche d'information. Tâche prometteuse pour une recherche plus ciblée et plus efficace.

Remerciements

Je tiens à saluer ici les personnes qui, de près ou de loin, ont contribué à la concrétisation de ce travail de thèse de doctorat. J'ai peut-être oublié certains noms qui ont été là pendant ces 4 ans de travail rigoureux, mais j'ai laissé au hasard de ma mémoire, plus impressionné par les événements récents, répétés, ou chargés d'émotions, le soin de retrouver ses personnes.

Je tiens tout d'abord à remercier M. Mohand BOUGHANEM pour avoir encadré et dirigé ma thèse, pour m'avoir fait une place au sein de l'équipe SIG-RFI. Je le remercie pour m'avoir soutenu et appuyé tout au long de ma thèse. Ses précieux conseils, son exigence et ses commentaires ont permis d'améliorer grandement la qualité de mes travaux et de ce mémoire.

Je tiens à remercier mon co-directeur de thèse, Monsieur Henri Prade, pour ces précieuses conseils et idées. Je le remercie pour sa collaboration et son encouragement.

Je tiens à exprimer ma profonde gratitude à M. Claude CHRISMENT, qui m'a fait l'honneur de présider le jury de thèse de doctorat, pour l'intérêt et le soutien chaleureux dont il a toujours fait preuve.

Je remercie Mme Gabriella PASI et Mme Sylvie CALABRETTO qui ont accepté d'être mes rapporteurs, pour l'honneur qu'ils me font en participant au jury. Je remercie également M. David Hawking d'avoir accepté d'examiner mon travail et de faire partie de Jury. Les commentaires et les questions de ces personnalités scientifiques, tant sur la forme du mémoire que sur son fond, ont contribué à améliore le document.

Mes remerciements vont de même à tous les membres de l'équipe SIG de l'IRIT pour leur accueil chaleureux et leur gentillesse.

Je remercie mes amis de l'équipe pour leur présence, leur aide et leur collaboration. Plus particulièrement, je tiens à remercier Mariam, Ordia, Mouna, Saad et Anas.

Merci à mes amis Aref, Chafic, Omar, bachar et Imad pour leur encouragement continu et leur fidélité.

Mes grands remerciements à mon père Walid, ma mère Zaïda qui constituent ma raison d'être et ma force pour continuer mes études. Mes sœurs Nibal, Nour et Rima et mon beau-frère Mony qui m'ont supporté et encouragé. Merci également à ma grande famille, mes oncles, mes tantes et tous mes cousins pour leurs encouragements et leur présence. Merci pour mes beaux parents Salman, Lamia et mes frères Louaye et Liwae pour leurs encouragements. Merci pour vous tous, vous n'avez jamais cessé de croire en moi pendant toutes mes années d'études.

Je remercie de tout mon cœur mon amour, Assala. Tu as été toujours présente dans les moments difficiles, tu m'as encouragé à atteindre mes buts, tu m'as supporté, tu as toujours resté à mes côtés. Tu es mon enthousiasme, mon bonheur, ma joie et tous ce qui est beau dans ma vie et tu restes pour toujours dans mon cœur.

Enfin, je close ces remerciements en dédiant cette thèse de doctorat à ma chérie Assala, mes parents, mes beaux parents, ma grande famille et à mes amis que j'ai eu la chance de les avoir à mes côtés et qui m'ont soutenu tout au long de ces années de travail.

Table des matières

Introduction Générale..... 15

PARTIE I : ETAT DE L'ART..... 19

Chapitre 1 : Recherche d'information, modèles et concepts de base..... 21

1.1 Introduction 21
1.2 Concepts de base de la RI..... 21
1.3 Système de recherche d'information (SRI) 22
1.3.1 Indexation..... 23
1.3.2 Pondération de termes 25
1.3.3 Appariement Requête-document 27
1.4 Les modèles de RI 28
1.4.1 Modèle booléen 28
1.4.2 Extension du modèle booléen..... 29
1.4.3 Modèle vectoriel..... 29
1.4.4 Les modèles probabilistes..... 30
1.4.4.1 Binary Independent Model..... 30
1.4.4.2 Modèle BM25..... 31
1.4.4.3 Modèle statistique de langue 32
1.5 Fonctionnalité additive à un SRI 33
1.6 Evaluation des SRI 34
1.6.1 Protocole d'évaluation..... 34
1.6.2 Les mesures d'évaluations..... 35
1.6.2.1 Les mesures de rappel précision..... 35
1.6.2.2 La courbe précision-rappel..... 35
1.6.3 La R-précision et la MAP (Mean Average Precision)..... 36
1.6.4 Présentation des campagnes d'évaluation des systèmes de recherche d'informations 37
1.6.4.1 La campagne d'évaluation TREC..... 37
1.6.4.2 Autres campagnes d'évaluations 38
1.6.4.3 Limites des campagnes d'évaluation des systèmes de recherches d'informations 39
1.6.5 Les outils d'évaluations..... 39
1.7 Conclusion..... 39

Chapitre 2 Représentation des textes 41

2.1 Introduction 41

2.2 Différentes formes qui peuvent avoir un descripteur 41
2.2.1 Unité représentative d'un descripteur 41
2.2.2 Approches statistiques 43
2.2.3 Approche syntaxique 43
2.2.4 Approches hybrides ou mixtes 44
2.2.5 Approches sémantiques 44
2.2.6 Approches conceptuelles 45
2.2.6.1 Méthode de désambiguïsation proposée par Baziz 46
2.3 WordNet, ressources externes pour l'extraction des descripteurs 47
2.3.1 Synset WordNet 48
2.3.2 Relations WordNet 48
2.3.3 Limites de WordNet 49
2.3.4 WordNet pour d'autres langues que le français 50
2.4 Modèle de représentation de l'information 50
2.4.1 Représentation par clusters 50
2.4.1.1 Représentation par cohésion lexical calculée par des relations extraites d'un thésaurus 52
2.4.2 Représentation sous forme de réseaux sémantiques 53
2.4.3 Représentation sous forme de graphes conceptuels 54
2.4.4 Autres modèles de représentations des connaissances 55
2.5 Résumé d'un texte ou d'un document 56
2.5.1 Les approches classiques 58
2.5.2 Les approches par apprentissage 59
2.5.3 Les approches basées sur le centroïde 60
2.5.4 Les approches exploitant la structure rhétorique 60
2.5.5 Les approches basées sur les graphes 61
2.6 Conclusion 61

PARTIE II : CONTRIBUTION 63

Chapitre 3 Extraction de termes significatifs d'un document 65

3.1 Introduction 65
3.2 Extraction de termes significatifs d'un texte 65
3.2.1 Extraction des concepts 66
3.2.2 Grouper les termes reliés 66
3.2.3 Facteurs pour caractériser les clusters et leurs concepts 68
3.2.3.1 Centralité d'un concept-terme 68
3.2.3.2 Fréquence d'un concept-terme 68
3.2.3.3 Fréquence conceptuelle 68

3.2.3.4 Spécificité 69
3.2.3.5 Taille du cluster 69
3.2.3.6 Poids du cluster 69
3.3 Conclusion 69

Chapitre 4 Vue sémantique d'un texte 71

4.1 Introduction 71
4.2 Les ensembles flous 71
4.3 Extraction d'une vue sémantique d'un texte 73
4.3.1 Sélection progressive des termes 73
4.3.2 Selection des mots à partir des clusters 74
4.3.3 Algorithme général 76
4.4 Expérimentation et résultats 76
4.4.1 Evaluation de l'extraction des termes significatifs 77
4.4.1.1 Collection de documents 77
4.4.1.2 Jugement de pertinence 78
4.4.1.3 Déroulement des expérimentations : 78
4.4.1.4 Analyse des résultats et discussion 80
4.4.2 Extraction de phrases significatives 81
4.4.2.1 Collection de documents 82
4.4.2.2 Jugement de la pertinence 83
4.4.2.3 Déroulement de l'experimentation 83
4.4.2.4 Analyse des résultats et discussion 84
4.5 Conclusion 87

Chapitre 5 Nouveaux facteurs pour la Recherche d'Information 89

5.1 Introduction 89
5.2 Désambiguïsation basée sur la centralité 89
5.3 La fonction c×f×s basé sur *c*, *f* et *S* pour la RI 90
5.4 Expérimentations et résultats 91
5.4.1 Collections de test TREC 91
5.4.2 Protocole d'évaluation 92
5.4.2.1 Baseline 92
5.4.2.2 Mesures d'évaluations 93
5.4.3 Les étapes d'expérimentations 93
5.4.4 Configurations utilisées 94
5.5 Meilleures valeurs de α, β *et* γ de notre fonction d'appariement 95
5.6 Optimiser les paramètres de « combinaison-paramètre(x,y,z) » 96

5.6.1 Evaluation des techniques de désambiguïsation des concepts requête et document . 96
5.6.2 Prise en compte des multi-termes vs terme simple 97
5.6.3 Evaluation de notre modèle d'appariement avec deux approches 98
5.6.3.1 Comparaison de notre modèle d'appariement à BM25 98
5.6.3.2 Comparaison de notre approche à celle de Baziz [Baziz, 2005] 100
5.7 Influence de la centralité et de la spécificité sur la pertinence 100
5.8 Impact des relations sur la centralité et la fréquence conceptuelle 101
5.8.1 Impact des relations WordNet pour représenter la centralité 102
5.8.2 Impact des relations WordNet pour représenter la fréquence conceptuelle 103
5.9 L'efficacité de la centralité et de la fréquence conceptuelle en RI 106
5.10 Conclusion 108

Conclusion Générale 109

REFERENCES BIBLIOGRAPHIQUES 113

Bibliographie 115

PARTIE III : Annexes 135

LISTE DES FIGURES

Figure 1: Processus général de recherche d'information (Baziz, 2005) 23
Figure 2: Courbe de distribution de mots selon la loi de Zipf 26
Figure 3: La correspondance entre l'informativité et la fréquence 27
Figure 4: Courbe de précision-rappel 36
Figure 5: Résultats de clustering pour grouper les carrés par leurs couleurs 51
Figure 6: Chaîne lexical d'un simple texte 52
Figure 7: Exemple de réseau sémantique 54
Figure 8: Le graphe conceptuel "une automobile est composée d'un moteur [Roussey, 2001] 55
Figure 9: Exemple sur les nuages des mots clés 57
Figure 10: Snippet Google 58
Figure 11: Exemple de création des clusters 67
Figure 12: Ensemble flou trapézoïdal 72
Figure 13: Partition floue des valeurs de centralité 72
Figure 14: Exemple d'ensemble flou assez-grand 73
Figure 15: Liste des sections internet 77
Figure 16: Comparaison graphique de la précision pour les différentes relations sur la collection TREC1 103
Figure 17: Comparaison graphique de la précision pour les différentes relations sur la collection TREC7 103
Figure 18: Exemple de relations entres cadres 137
Figure 19: Extrait de la base de connaissance DBpédia 140
Figure 20: Trois structures topologiques 141
Figure 21: Exemple sur les noeuds intermediaire 143
Figure 22: Comparaison graphique de la précision pour les différentes relations sur la collection TREC1 148
Figure 23: Comparaison graphique de la précision pour les différentes relations sur la collection TREC7 149
Figure 24: Comparaison graphique de la précision pour les différentes relations sur la collection TREC1 pour la fonction BM25 150
Figure 25: Comparaison graphique de la précision pour les différentes relations sur la collection TREC1 pour le modèle de langue 151
Figure 26: Comparaison graphique de la précision pour les différentes relations sur la collection TREC7 pour la fonction BM25 153
Figure 27: Comparaison graphique de la précision pour les différentes relations sur la collection TREC1 pour le modèle de langue 154

LISTE DES TABLEAUX

Tableau 1: Extrait de résultats obtenus avec quelques fonctions d'agrégation sur le document "internet" et sur les ses sous-sections 80
Tableau 2: Liens vers les documents de la collection 83
Tableau 3: Les phrases extraites manuellement comparant à celles extraites automatiquement 84
Tableau 4: Résultats obtenus pour l'ensemble de documents 86
Tableau 5: Différentes variantes utilisés dans les expérimentations 95
Tableau 6: Impact des différents paramètres 96
Tableau 7: Comparison des résultats obtenus par les différentes méthodes de désambiguïsation de concept document 97
Tableau 8: Impact des concept multi-termes 98
Tableau 9: Comparer nos résultats à celles obtenus par BM25 99
Tableau 10: Comparaison entre la precision obtenue par nos résultats et celle obtenue par une méthode conceptuelle 100
Tableau 11: Les dix premiers documents pertinents selectionnés selon BM25 101
Tableau 12: Les dix premiers documents pertinents retirés selon BM25 101
Tableau 13: Impact des relations de Cf sur les top-10 documents obtenus sur différentes fonctions et pour les deux collections TREC1 et TREC7 105
Tableau 14: Résultats de c^2*f*s 106
Tableau 15: Résultats d'Okapi-BM25 basé sur Cf 107
Tableau 16: Résultats du lissage de Dirichlet basé sur Cf 107
Tableau 17: Liste des étiquettes effectives 138
Tableau 18: Degré d'opinion affecté pour les 3 sens de l'adjectif "estimable" [Chaumartin, 2007b] . 139
Tableau 19: Les résultats obtenus par la fonction d'agrégation $\max(\mu_{\text{tf-grandsuffisament}}, \min(\mu_{\text{s-grand}}, \mu_{\text{c-grand}}))$ 146
Tableau 20: Impact des différentes relations sur la fonction c^2*f*s pour la collection TREC1 147
Tableau 21: Impact des différentes relations sur la fonction c^2*f*s pour la collection TREC1 147
Tableau 22: Impact des différentes relations sur la fonction c^2*f*s pour la collection TREC7 148
Tableau 23: Impact des différentes relations sur la fonction c^2*f*s pour la collection TREC7 149
Tableau 24: Impact des différentes relations sur la fonction BM25 pour la collection TREC1 150
Tableau 25: Impact des différentes relations sur la fonction BM25 pour la collection TREC1 150
Tableau 26: Impact des différentes relations sur la fonction du Modèle de Langue pour la collection TREC1 151
Tableau 27: Impact des différentes relations sur la fonction du Modèle de Langue pour la collection TREC1 151
Tableau 28: Impact des différentes relations sur la fonction BM25 pour la collection TREC7 153
Tableau 29: Impact des différentes relations sur la fonction BM25 pour la collection TREC7 153
Tableau 30: Impact des différentes relations sur la fonction du Modèle de Langue pour la collection TREC7 154
Tableau 31: Impact des différentes relations sur la fonction du Modèle de Langue pour la collection TREC7 154

Introduction Générale

Contexte

Nos travaux se situent dans le contexte des systèmes de recherche d'information : l'objectif d'un système de recherche d'information est de sélectionner à partir d'un ensemble de documents les informations pertinentes susceptibles de répondre à la requête. Afin d'y parvenir, une des tâches clés d'un système de recherche d'information est l'extraction des termes pouvant représenter le document. Ce processus est appelé indexation, il permet de construire un ensemble d'éléments clés qui caractérise le contenu d'un document, ces éléments peuvent être utilisés dans différentes tâches de RI, la recherche adhoc, la classification, le filtrage, le résumé, etc. Ces éléments sont souvent des mots simples ou groupe de mots extraits automatiquement ou manuellement, directement du document ou peuvent être sélectionnés à partir d'une ressource externe comme, les lexiques qui sont des listes de mots clés, des ontologies ou des thésaurus.

Une fois ces éléments extraits, des fonctions permettant de les caractériser leur sont associées, ce sont souvent les fréquences du terme dans le document, sa position dans le texte, sa fréquence inverse (IDF) et la longueur du document. Ces facteurs sont ensuite combinés pour répondre à différentes tâches de RI, soit pour mesurer la pertinence d'un document vis-à-vis d'une requête selon un modèle donné, ou pour mesurer la pertinence d'une phrase, ou d'un paragraphe pour résumer le document, ou encore la pertinence d'un document pour être classé dans une catégorie donnée.

Les travaux présentés dans ce mémoire abordent deux questions, la première concerne l'extraction des termes significatifs dans un document et leur caractérisation. La seconde pointe sur la prise en compte de ces termes et des facteurs associés dans deux tâches distinctes : l'extraction d'une vue sémantique d'un document, une suite de résumé du texte formé de mots importants, et dans la tâche de recherche d'information (adhoc) classique.

Problématique

L'introduction rappelle plusieurs points. Le premier concerne le type d'éléments (représentation par mots clés, groupes de mots, concepts), à considérer que représenter le document, c'est en effet une question récurrente en RI, toujours d'actualité. La représentation basée sur les mots-clés atteint certaines limites. L'intervention d'une représentation basé-concepts devient nécessaire pour résoudre le problème de la polysémie des unités langagières que représentent les mots. Dans cette représentation, les documents et les requêtes sont représentés par un ensemble de descripteurs (ou concepts) appartenant à une terminologie prédéfinie avant l'indexation. Ces terminologies peuvent être un simple vocabulaire, un thésaurus ou une ontologie fournissant des relations entre les concepts. La couverture de ces terminologies pourra être limitée à un domaine (connaissances techniques par exemple), associer différents types de descripteurs (e.g. de structure et sémantiques), voire correspondre

à un vocabulaire général (type WordNet). Cette représentation fait apparaître de nouveaux problèmes pour rattacher un terme à un meilleur concept qui peut le décrire, cela est dû au fait de la synonymie et de la polysémie présente dans le langage naturel. Des algorithmes de désambiguïsation seront nécessaires pour corriger les concepts retirés. Une problématique que nous traitons dans ce mémoire consiste à proposer une méthode de désambiguïsation permettant de trouver le meilleur concept représentatif d'un terme du document.

Le second point concerne la vue sémantique d'un texte pour faciliter sa compréhension ou encore avoir une idée de son contenu sans forcément le lire. La vue sémantique décrit les informations représentatives du contenu du document, c'est-à-dire des informations auxquelles un utilisateur est susceptible de faire référence lors d'une recherche. Elle peut être composée de différents éléments, tels que liste de mot clés, ensemble de phrases ou de paragraphes extraits du texte, etc. La problématique reliée aux vues sémantiques concerne la manière d'extraire et de représenter ces éléments. En fait, la plupart des approches existantes utilisent des méthodes purement statistiques afin de représenter le contenu des textes. Cependant, il semblerait qu'une meilleure vue puisse en être obtenue en allant au-delà du simple décompte, en prenant en compte des groupes de mots appartenant à un même champ lexical que nous nommons « cluster ». De plus, les mots d'un « cluster » peuvent être plus ou moins spécifiques, et ainsi contribuer différemment à la perception immédiate par un lecteur des thématiques d'un texte. Dans ce mémoire nous proposons un modèle de représentation de la vue sémantique d'un document indépendamment de toute requête utilisateur.

La troisième concerne les facteurs *tf*, *idf*, *taille du document* utilisés dans les modèles de RI pour pondérer les termes d'un document. D'une part, ces facteurs sont reliés à la présence des termes dans un document, un terme qui n'apparait pas dans le document ne sera affecté d'aucun poids. Ainsi, un document pertinent ne contenant pas les termes de la requête ne sera jamais sélectionné dans la liste des documents pertinents. D'autre part, leurs calculs se font indépendamment du sens des termes. Ainsi par exemple les documents retournés en réponse à une requête contenant le terme « *data* » contiennent exactement ce terme malgré que des documents contenant le terme « *information* » soient aussi pertinents pour la même requête. Pour résoudre ces problèmes nous avons introduit de nouveaux facteurs permettant à notre sens, de mieux caractériser la dimension sémantique du contenu d'un document. Ces facteurs vont au-delà du simple comptage des termes, ils tentent de mesurer combien un terme donné est relié (sémantiquement) aux autres termes d'un document. Ces facteurs sont la « *centralité* », la « *fréquence conceptuelle* » et la « *spécificité* ».

Contribution : Nouveaux paramètres à base ontologique pour la recherche d'information

Nous avons proposé une méthode de représentation de document par « cluster », où nous groupons dans une même classe les termes se référant à un même concept (un thème). Pour ce faire, dans un premier temps nous avons attaché à chaque terme extrait du document un concept qui lui correspond dans une ontologie. Ensuite, nous avons groupé les concepts qui ont une relation sémantique dans l'ontologie dans des groupes que nous avons nommés « clusters ». A partir de cette présentation par « cluster », nous avons pu définir nos facteurs sémantiques. Ces facteurs sont estimés de deux manières soit par le calcul de la « *centralité* » d'un concept, soit par le calcul de la « *fréquence conceptuelle* » d'un concept. Avec la centralité $c_{t,d}$ d'un concept t est estimée par le nombre de concept du document d qui sont en

relations sémantiques avec *t*. La fréquence conceptuelle $Cf_{t,d}$ d'un concept *t* est estimée par la somme des fréquences des concepts qui sont en relation direct avec *t* dans le document *d*. En plus de ces deux facteurs, un autre facteur « *la spécificité* » qui estime combien un terme d'une requête est spécifique ou générique dans un contexte spécifique. La spécificité S_t d'un terme *t* est estimée au moyen de sa "profondeur" dans l'arbre conceptuel induit par la relation "est-un" dans l'ontologie. Ces facteurs combinés avec les facteurs classiques *tf*, *idf* et *ld* (la longueur du document) sont utilisés dans trois méthodes différentes. La première méthode porte sur la représentation du contenu sémantique d'un texte. La deuxième correspond à la problématique (déjà évoquée dans la partie « problématique ») de désambiguïsation des concepts représentatifs d'un document où nous avons proposé une nouvelle méthode de désambiguïsation basée sur la centralité. Dans la troisième méthode, nous intégrons ces différents facteurs dans la recherche d'information pour pondérer un document vis-à-vis d'une requête.

Plus précisément, dans la première méthode nous avons proposé une technique d'extraction du contenu sémantique d'un texte indépendamment de toute requête. Cette vue permet de représenter le document par groupes flous de concepts pondérés répartis sur plusieurs niveaux. Avec, les mots représentatifs à chaque niveau sont extraits du texte en se basant sur des fonctions d'agrégation de nos facteurs sémantiques et des facteurs classiques. Ainsi, chaque niveau fournit plus de détails sur le contenu du document. Tout d'abord, les mots les plus importants et significatifs sont identifiés à partir des clusters dans les niveaux supérieurs. Plus on descend de niveau moins les termes seront significatifs. Suite à cette représentation multi-niveau d'un texte par groupes de mots significatifs, quelques phrases contenant un maximum de ces mots sont extraites du texte, et proposées comme représentatives de son contenu.

Dans la deuxième méthode, nous avons proposé une technique de désambiguïsation des concepts basée sur la centralité. En fait, un terme extrait d'un document peut être attaché à différents concepts dans l'ontologie, dont chaque concept représente un sens de terme. Le but de la méthode proposée est de trouver le concept représentant le meilleur sens du terme dans son contexte d'apparition. Dans ce but, nous avons calculé la *centralité* de chaque concept candidat pour représenter le terme, ainsi nous n'avons gardé que celui qui a la plus grande centralité.

Dans la troisième méthode, nous avons utilisé nos facteurs sémantiques pour pondérer un document vis-à-vis d'une requête. D'une part, nous avons proposé notre propre fonction de pondération. D'autre part, nous avons intégré nos paramètres dans des fonctions de pondérations connus tels que la fonction BM25 et une fonction connu du modèle de langue qui correspond précisément au lissage d'indépendance de Dirichlet. Dans cette méthode nous avons mené une étude sur les différentes relations qui influencent le calcul de la *centralité* et la *fréquence conceptuelle*.

Plan de la thèse

La thèse est constituée, en plus de cette d'introduction, de deux parties. La première partie comporte deux chapitres sur l'état de l'art en lien avec le cadre de notre thèse. La deuxième partie, que nous avons subdivisée en trois chapitres, regroupe les différents aspects de notre contribution.

Plus précisément, la première partie concerne les notions et concepts de base de la RI ainsi que les principaux modèles existants, et dresse un état de l'art sur l'indexation sémantique et conceptuelle où les principaux travaux concernant l'utilisation de la sémantique et des concepts en RI sont détaillés. Le chapitre 1 présente l'architecture générale d'un système de recherche d'information et les concepts de base de RI. Par la suite, on décrit les principaux modèles qui sont à la base de la majorité des systèmes actuels. Enfin, les principales mesures de pertinence utilisées en RI ainsi que les principaux cadres d'évaluation des systèmes expérimentaux sont décrits.

Le chapitre 2 est consacré à l'état de l'art sur l'extraction des descripteurs à partir des documents. Ce processus, nommé processus d'indexation, permet de trouver les éléments significatifs, souvent des mots clés, qui représentent le contenu d'un document. Les descripteurs ainsi extraits peuvent avoir différents formes (mots clés, mot-sens, concepts) permettent et peuvent être utilisé pour représenter le document sous forme *nuages de mots-clés* (*Tag Cloud* ou *Word Cloud*).

La deuxième partie de la thèse comprend trois chapitres. Dans le chapitre 3, la méthode de représentation sémantique d'un document par cluster des concepts est détaillée. Nous y définissons les différents paramètres sémantiques tels que la taille du cluster, son *poids* qui est la somme de fréquence des mots qui le composent, mais aussi la *spécificité* de ces mots, ou leur *centralité* dans les "clusters", et la *fréquence conceptuelle* etc. Cette représentation ainsi que les différents paramètres sont ensuite utilisés dans des directions différentes que nous les présentons dans les chapitres 4 et 5.

Dans le chapitre 4, nous détaillons la méthode d'extraction du vues sémantiques d'un texte. Nous utilisons la représentation par cluster pour représenter le contenu sémantique d'un texte sous forme de groupes thématiques pondérés. Nous présentons la méthode d'extraction de point de vue sémantique multi-niveau du contenu du texte ainsi que l'extraction des phrases significatives d'un texte. L'intérêt de cette méthode est testé dans des expérimentations.

Dans le chapitre 5, nous représentons la méthode de désambiguïsation des concepts ainsi que la méthode de pondération d'un document vis-à-vis d'une requête en combinant les paramètres sémantiques que nous avons définis. Les méthodes proposées sont suivies d'une liste des expérimentations montrant leur efficacité.

Notre terminons par une conclusion générale qui représente un bilan sur les travaux réalisés. Les différentes méthodes ainsi que les résultats sont représentés, ensuite les perspectives proposées envisageables pour ces travaux dans le cadre de l'utilisation des paramètres sémantiques dans les systèmes de recherche d'informations.

PARTIE I : **ETAT DE L'ART**

Chapitre 1: Recherche d'information, modèles et concepts de base

1.1 Introduction

Le domaine de recherche d'information fut créé à cause de besoin de trouver parmi les documents électroniques disponibles, l'information qui correspond à nos besoins en un minimum de temps. Le terme de « recherche d'information » (information retrieval) fut donné par Calvin N. Mooers en 1948 pour la première fois lors de son travail sur son mémoire de Maîtrise [Mooers, 1948]. Les systèmes de recherche d'information (SRI) ont été conçus dans le but de retrouver des informations dans des collections de documents.

Dans ce chapitre, nous présentons les concepts de base d'un système de recherche d'information. Nous décrivons ensuite les modèles de recherche d'information (RI) les plus populaires. Nous terminons ce chapitre par les processus d'évaluation mis en place pour mesurer l'efficacité de ces systèmes et les différents projets qui ont été à la base de l'évolution de ce domaine.

1.2 Concepts de base de la RI

Selon [salton, 1971b] la recherche d'information est l'ensemble des techniques permettant de sélectionner à partir d'une collection de documents, ceux qui sont susceptibles de répondre au besoin de l'utilisateur exprimé via une requête. A travers cette définition, nous retenons 3 concepts clés : le document, le besoin et la pertinence.

Document

Le document est constitué par un texte, un morceau de texte, une image, une bande de vidéo etc., qui peut être retourné en réponse à une requête/ besoin en informatique d'un utilisateur. L'ensemble des documents sur lequel porte une recherche forme la collection de documents.

Besoin en info et requête

Le besoin en information d'un utilisateur peut avoir différents types : *besoin vérificatif* où l'utilisateur cherche une donnée particulière et sait souvent comment y accéder (exemple l'utilisateur cherche un document avec une adresse web), *besoin thématique* connu où l'utilisateur cherche à trouver une nouvelle information concernant un sujet ou un domaine connu, enfin *un besoin thématique inconnu*, l'utilisateur dans ce cas cherche de nouveaux concepts ou de nouvelles relations hors des sujets ou domaines qui lui sont familiers.

La requête formule le besoin de l'utilisateur sous forme d'un ensemble de mots exprimé en langage naturel, booléen ou graphique. La requête est soumise à un moteur de recherche pour une recherche documentaire donnée.

Pertinence

La notion de pertinence est un critère principal pour l'évaluation des systèmes de recherche d'information. De nombreuses définitions possibles, selon [Saracevic, 1970] *la pertinence peut être vue comme une mesure de l'informativité du document à la requête*, *un degré de relation (chevauchement, etc.) entre le document et la requête* ou *un degré de surprise qu'apporte un document vis-à-vis du besoin de l'utilisateur* etc.

La pertinence est subjective, c'est-à-dire elle dépend de l'utilisateur. Les études menées autour de la notion de pertinence [Barry, 1994] [Borlund, 2003] montrent qu'elle est définie

par un ensemble de critères et des préférences qui varient selon les utilisateurs (*le contenu informationnel des documents*, *le niveau d'expertise et de connaissances de l'utilisateur, des informations liées à l'environnement, les sources des documents, etc.*). Ces critères sont des facteurs qui déterminent la pertinence accordée à l'information retrouvée par l'utilisateur dans un contexte de recherche précis.

1.3 Système de recherche d'information (SRI)

Comme illustré dans la Figure 1, un SRI intègre deux fonctions principales, représentées schématiquement par le processus U de recherche d'information [Belkin et al, 1992] : *indexation document* (*resp. requête*) et *appariement document-requête*. Plus précisément

- Processus d'indexation : consiste à extraire le descripteur à partir de la requête ou des documents. Le descripteur est une liste des termes significatifs qui doit couvrir au mieux le contenu sémantique d'un document ou d'une requête. L'ensemble des termes reconnus par le SRI est rangé dans une structure appelée dictionnaire constituant le langage d'indexation.
- Processus de recherche : il permet de sélectionner l'ensemble des documents potentiellement pertinents pour une requête. Le cœur de ce processus est une fonction de correspondance entre la requête et les documents par exploitation de l'index, ensuite présentation des résultats. Cette fonction de correspondance calcule un score de pertinence *RSV(Q,D)* (*Retrieval Status Value*) entre la requête indexée *Q* et les descripteurs du document *D*. Différents modèles de RI ont été proposés pour mesurer ces scores. Ces modèles sont au centre des travaux de recherche dans le domaine de la RI, ils seront détaillés dans la section 1.4. Les documents retournés généralement sont triés par ordre de leurs pertinences, du plus pertinent au moins pertinent (c.à.d. par ordre de valeur de correspondance décroissante)

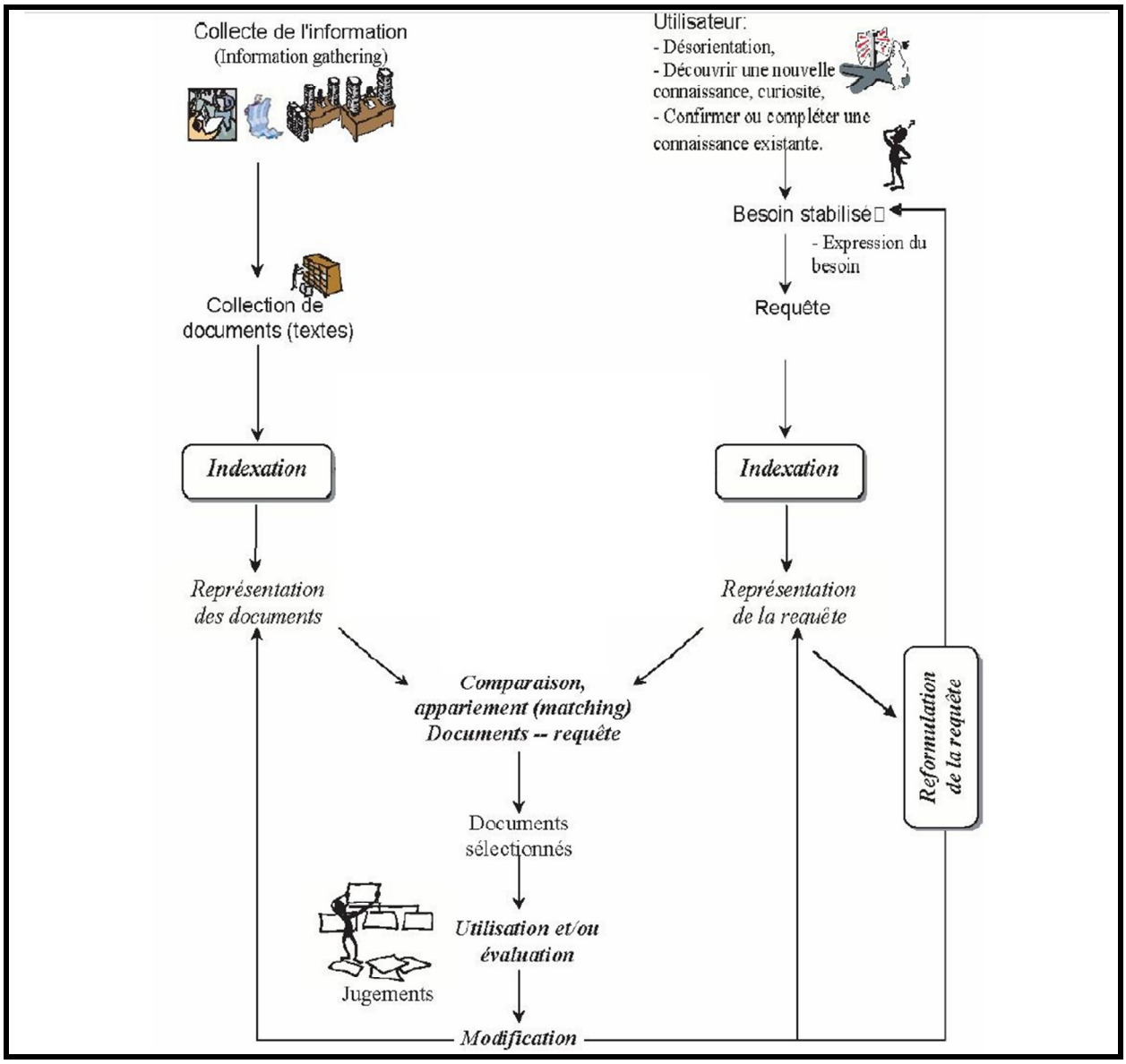


Figure 1: Processus général de recherche d'information (Baziz, 2005)

1.3.1 Indexation

La tâche principale d'un processus d'indexation est de transformer les documents (ou requêtes) en substituts ou descripteurs capables de représenter leurs contenus [Salton, 1971b][Sparck Jones, 1979][Van Rijsbergen, 1979]. Dans une indexation classique, les descripteurs d'un document peuvent être des termes simples ou des termes composés. Dans des systèmes qui prennent en compte le sens des termes et les différents types de relations sémantiques dans un document, ces descripteurs seront des mots-clés (concepts) associés à des entrées dans un vocabulaire contrôlé, thésaurus [Salton et al, 1968] [Sparck Jones, 1986], ontologie (exemple PenMan, Cyc [Sowa, 1984] et WordNet [Miller, 1995], hiérarchie de concepts, etc.). Les descripteurs sont ensuite stockés dans une structure particulière appelée fichier inverse. Dans ce fichier, nous trouvons pour chacun de ces termes, la liste des références de chaque document le contenant. Par référence on entend identifiant, c'est-à-dire un moyen de retrouver de façon non ambigüe des documents ou un document ou une partie de document où le terme apparaît.

Pour améliorer l'indexation plusieurs approches d'amélioration de la représentation d'un document ont été proposées, citons :

Stoplists : Dans le but de représenter le document par les termes qui représentent mieux son contenu, des approches à base fréquentielle ont été employées comme nous les détaillerons dans la section suivante. Reste toujours certains mots qui sont fréquents dans le document mais ils ne lui donnent pas une valeur discriminante. Ces mots sont souvent des prépositions (e.g. "de", "à"), prénom ("aucun", "tout", "on"), certains adverbes ("ailleurs", "maintenant"), adjectifs ("certain", "possible"), etc. Afin d'éliminer ces mots et améliorer la représentation on utilise une liste appelée stoplist (ou liste des mots vides).

Certains mots inclus dans cette liste ne sont pas nécessairement vides de sens (ça dépend du domaine). Ils ne sont pas vides de sens en linguistiques, mais leur sens importe très peu pour des besoins de RI.

Le traitement lié à une stoplist est très simple. Quand on rencontre un mot dans un texte, on doit d'abord examiner s'il apparaît dans cette liste. Si oui, on ne le considère pas comme un mot à insérer dans l'index.

Lemmatisation : Les mots (lemmes) d'une langue utilisent plusieurs formes en fonction de leur genre (masculin ou féminin), leur nombre (un ou plusieurs), leur personne (moi, toi, eux, ...), leur mode (indicatif, impératif, ...) donnant ainsi naissance à plusieurs formes pour un même lemme. La lemmatisation désigne l'analyse lexicale du contenu d'un texte regroupant les mots d'une même famille. Chacun des mots d'un contenu se trouve ainsi réduit en une entité appelée lemme (forme canonique). La lemmatisation regroupe les différentes formes que peut revêtir un mot, soit : le nom, le pluriel, le verbe à l'infinitif, etc.

Plusieurs approches de lemmatisation ont été proposées : Porter [Porter, 1980] a créé un algorithme qui élimine les terminaisons des mots en anglais en 5 grandes étapes: la première étape tente de transformer le pluriel en singulier. Les étapes subséquentes essayent d'éliminer au fur et à mesure les dérivations. Cet algorithme transforme parfois deux mots différents en une même forme. Par exemple, derivate/derive, activate/active. Cependant, pour la plupart, la transformation semble raisonnable. Maintenant, la plupart de procédures de lemmatisation l'utilise, ou utilise une de ces variantes.

Autre algorithme de lemmatisation utilise un dictionnaire pour savoir si une séquence de lettre à la fin correspond à une terminaison [Savoy, 1993], il suffit de faire une élimination ou une transformation tentative, et de voir si la forme obtenue existe dans le dictionnaire.

D'autre modèle de lemmatisation, utilise un taggeur (ou un analyseur de catégorie) automatique. Dans ces approches, afin de déterminer la catégorie d'un mot on utilise un modèle probabiliste. Ce modèle détermine la probabilité qu'un mot soit présent dans une catégorie selon sa forme, et selon les mots qui l'entourent.

Le processus d'indexation dépend de plusieurs facteurs, en particulier la manière d'indexer : automatique ou manuelle, et la source utilisée pour représenter les descripteurs (vocabulaire contrôlé ou libre extrait directement du texte du document).

En RI, ils existent trois modes d'indexation: l'indexation manuelle, automatique ou semi-automatique.

- Indexation manuelle : les descripteurs d'un document sont choisis par un expert de domaine qu'il juge pertinents pour la description du contenu sémantique du document. Ce type d'indexation permet d'avoir un vocabulaire d'index contrôlé, ce qui permet d'accroître la consistance et la qualité de la représentation obtenue [Nie et al, 1999], par conséquent les documents retournés par un SRI en réponse à une

requête utilisateur sont précis [ren et al, 1999]. Mais l'augmentation du nombre de documents à indexer rend la tâche d'indexation manuelle difficile et coûteuse en temps. De plus cette indexation dépend des connaissances des spécialistes qui peuvent être limitées à cause de l'évolution de la langue (apparition de nouveaux mots).

- Indexation automatique : La première approche à l'indexation automatique KWIC ou Keyword in Context, fut introduite à International Conference on Scientific Information (ICSI) en 1958 par Luhn. Dans ce type d'indexation les descripteurs sont automatiquement extraits à partir du texte du document. La fréquence d'occurrence de mots a été utilisée comme le critère de sélection d'index. Les mots vides de sens (ceux de *stoplistes*) sont systématiquement éliminés.
- Indexation semi-automatique ou encore notée indexation supervisée : consiste à tirer profit des 2 types d'indexation manuelle et automatique, certains travaux [Jacquemin et al, 2002] proposent d'exposer les résultats de l'indexation automatique à un spécialiste de domaine. Le spécialiste utilise un vocabulaire contrôlé sous forme de thésaurus ou de base terminologique qui couvre le langage d'indexation.

1.3.2 Pondération de termes

La pondération est l'opération qui consiste à affecter un poids aux termes d'indexation et de recherche. Ce poids permet de préciser l'importance relative des mots représentés dans les documents par rapport à ceux identifiés dans la question, selon une échelle déterminée.

Différents facteurs basés sur l'occurrence statistique ont été utilisés pour la pondération des termes d'un document :

La fréquence tf (ou Term Frequency en anglais): *tf* quantifie la représentativité locale d'un terme dans le document. Cette valeur est fonction de la fréquence d'occurrence du terme dans le document. Cependant on s'aperçoit que les mots les plus fréquents sont des mots fonctionnels (ou mots vides). En français, les mots "de", "un", "les", etc. sont les plus fréquents, en anglais, ce sont "of", "the", etc. L'intérêt et la spécificité de la fréquence des mots ont été étudiés à la fin des années 48 par Zipf [Zipf, 1949]. Selon Zipf, les mots dans les documents ne s'organisent pas de manière aléatoire mais suivent une loi inversement proportionnelle à leur rang. Le rang d'un mot est sa position dans la liste décroissante des fréquences des mots du corpus. Formellement, cette loi s'exprime de la manière suivante :

$$Fréquence \times Rang \simeq Constante$$

Selon cette loi, la distribution des mots suit la courbe suivante (Figure 2):

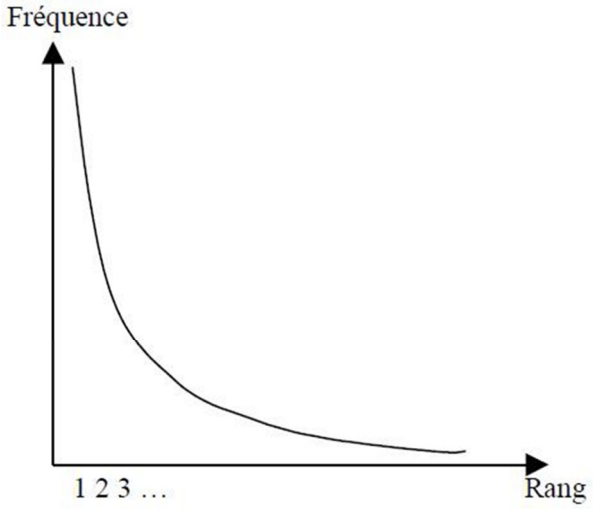


Figure 2: Courbe de distribution de mots selon la loi de Zipf

La fréquence de termes semble donc être un bon indicateur pour mesurer l'importance d'un terme. A ce titre, Luhn [Luhn, 1957] définit l'informativité d'un terme en se basant sur la loi de Zipf. L'informativité mesure la quantité de sens qu'un mot porte. Cette notion n'est pas définie très précisément en RI. Elle est utilisée seulement de façon intuitive. Cependant, on peut trouver son équivalent dans la théorie de l'information (par exemple, la théorie de Shannon qui est une théorie probabiliste permettant de quantifier le contenu moyen en information d'un ensemble de messages, dont le codage informatique satisfait une distribution statistique précise [Shannon, 1948]).

La correspondance entre l'informativité et la fréquence est illustrée sur la Figure 3. Ainsi, les mots ayant des fréquences entre les deux seuils Min et Max sont les mots qui ont l'informativité la plus élevée.

Les fonctions de calcul de *tf*:

- *tf* = 1 ou 0, dans le cas d'une fonction binaire qui rend 1 si le terme apparaît dans le document une ou plusieurs fois, et rend 0 si le terme n'apparait pas.
- *tf* = f(t,d), où f(t,d) est une fonction brute de calcul de la fréquence qui compte le nombre d'occurrence du terme *t* dans un document *d*.
- $tf = \alpha + \log(f(t,d) + \beta)$, dans le cas d'une fonction logarithmique, où α et β sont des constantes. Cette fonction a pour but d'atténuer les effets de larges différences entre les fréquences d'occurrence des termes dans le document.
- *tf* = f(t,d)/Max[f(t,d)], dans le cas d'une fonction normalisée, où Max[f(t,d)] est la fréquence maximale des termes dans d; Cette fonction permet de réduire les différences entre les valeurs associées aux termes du document.
- $tf = [f(t,d) * (K_1 + 1)] / [k_1 * ((1-b) + b * dl/avdl)]$, Robertson [Robertson, 1994a] [Robertson, 1994b] dans son modèle BM25 utilise une normalisation de la fréquence d'un terme *t* dans un document *d* en fonction du longueur de *d* noté par *dl* et la moyenne de longueur des documents dans la collection traitée *avdl*, où K_1 et *b* sont des constantes.

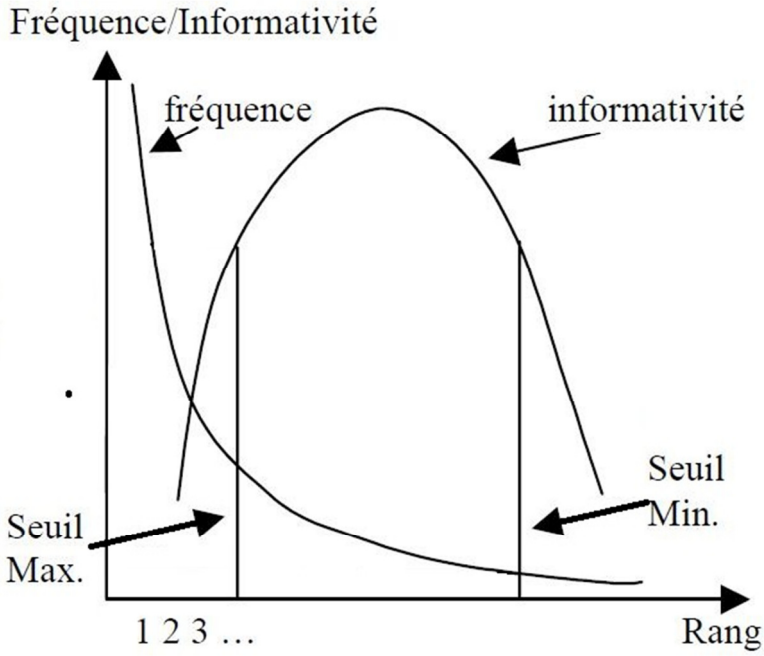


Figure 3: La correspondance entre l'informativité et la fréquence

La fréquence inverse idf (ou Inverse Term Frequency en anglais) : *idf*, quantifie la représentativité globale du terme dans une collection de documents. Cette valeur mesure si le terme est discriminant, les termes qui sont utilisés dans de nombreux documents sont moins discriminants que ceux qui apparaissent dans peu de documents. Un poids plus important est donc assigné aux termes qui apparaissent moins fréquemment dans la collection.

Citons par exemple une fonction de calcul d'idf :

idf = log(N/n) ou log((N-n)/N), où *N* désigne le nombre totale de documents dans la collection et *n* désigne df(t,C) (le nombre de document de la collection *C* contenant le terme *t*)

*La pondération basée sur tf*idf :* d'une manière générale, la majorité des méthodes de pondération [Robertson et al, 1997] [Singhal et al, 1997] [Sparck Jone, 1979] est construite par la combinaison de deux facteurs, *tf* "Term Frequency" et *idf* "Inverse of Document Frequency". La formule tf*idf assigne à un term *t* un poids dans un document *d* qui est : 1) le plus grand lorsque *t* apparait dans un petit nombre de document. 2) petit lorsque le terme a une petite occurrence dans le document. 3) le plus petit lorsque le terme apparait virtuellement dans tous les documents de la collection.

La formule finale de tf*idf est donc la multiplication de tf par idf. Par exemple :

Tf*idf = [f(t,d)/Max[f(t,d)]] * log(N/n)

1.3.3 Appariement Requête-document

Cette étape permet de comparer la représentation d'un document à celle de la requête. Ceci revient souvent à mesurer un score de correspondance (de pertinence) entre ces représentations souvent appelé RSV (Retrieval Status Value). Ce score prend en considération les descripteurs, ainsi que, leurs pondérations dans la représentation de la requête et la représentation du document.

Plus précisément, l'appariement requête-document dépend du modèle de RI utilisé. Un modèle de RI définit la manière dont la requête et les documents sont représentés, ainsi que, le modèle formel qui permet d'interpréter cette notion de pertinence. Plusieurs modèles ont été

proposés dans le domaine comme les modèles vectoriels, probabilistes, connexionnistes, inférentiels, qui seront détaillés par la section 1.4.

1.4 Les modèles de RI

Le rôle d'un modèle est fondamental dans un SRI. Il permet d'interpréter la notion de pertinence d'un document vis-à-vis une requête dans un cadre formel. Il fournit donc un cadre théorique pour la modélisation de cette mesure de pertinence. Plusieurs modèles, s'appuyant sur des cadres théoriques allant des ensembles aux probabilités en passant par l'algèbre ont été définis.

Le modèle booléen était le premier à être employé dans la RI, à cause de sa simplicité. Cependant, le manque de pondération dans ce modèle limite ses utilisations. Ainsi, des versions étendues de ce modèle ont été proposées, elles intègrent la pondération, exemple l'utilisation de la théorie des ensembles flous [Kraft et al, 1983]. Le modèle vectoriel est sans doute le modèle le plus souvent utilisé en RI. Sa popularité est due à sa capacité d'ordonner les documents retrouvés et à ses bonnes performances dans les tests.

Depuis le milieu des années 1970, C.J. van Rijsbergen [Van Rijsbergen, 1977], S. Robertson et K. Sparck Jones [Robertson et al, 1976] étaient parmi les premiers à proposer les modèles probabilistes. La performance de ces modèles apparaît dans les années 1990. Ensuite, les modèles de langue [Ponte et Croft, 1998] ont eu de succès en raison de leur simplicité théorsique.

Nous présentons dans la suite le principe de ces modèles : modèle booléen, modèle vectoriel et modèle probabiliste.

1.4.1 Modèle booléen

La stratégie de recherche booléenne était employée dans les premiers systèmes de gestion de bibliothèques. Il est basé sur la théorie des ensembles [Salton, 1971b]. Dans ce modèle chaque document *d* est représenté comme une conjonction logique de termes (non pondérés) qui le composent, par exemple :

$$d = \{\mathrm{t1}\,, \mathrm{t2}\,, \ldots, \mathrm{tn}\}$$

Une requête de l'utilisateur est formulée à l'aide d'une expression booléenne. La requête booléenne est représentée par peu des termes reliés par des opérateurs booléens :

$$q = (\mathrm{t1} \wedge \mathrm{t2}\,) \vee (\mathrm{t3} \wedge \neg \mathrm{t4})$$

La correspondance $RSV(D,Q_k)$, entre une requête $Q_k(q_{k1}, q_{k2},\ldots,q_{kn})$ composée des termes q_{ki} avec i ∈ [1, n] et une listes documents représentés par *D* est déterminée comme suit :

$RSV(D,q_{ki})= R(q_{ki}\,)$ *où* $R\,(q_{ki})$ *est l'ensemble de documents retrouvé par* q_{ki}

$RSV(D_j,q_{ki} \wedge q_{kj})=R(q_{ki}) \cap R(q_{kj})$

$RSV(D_j,q_{ki} \vee q_{kj})= R(q_{ki}) \cup R(q_{kj})$

L'intérêt du modèle booléen est qu'il permet de faire une recherche très restrictive et obtenir, pour un utilisateur expérimenté, une information exacte et spécifique.

Les inconvénients de ce modèle sont les suivants :

- Les réponses à une requête ne sont pas ordonnées puisque la correspondance entre un document et une requête est soit 1, soit 0. Il n'est pas possible de dire quel

document est mieux qu'un autre. Le problème devient plus difficile si les documents qui répondent aux critères de la requête sont nombreux.

- L'expression d'une requête nécessite une connaissance des opérateurs booléens, ceci n'est pas une tâche simple pour tous les usagers.

Avec tous ces inconvénients, le modèle booléen standard reste toujours utilisé. Des extensions ont été proposées pour tenir compte de l'importance des termes dans les documents. On y trouve par exemple le modèle booléen étendu [Salton et al, 1983] ou encore un modèle basé sur les ensembles flous [Zadeh, 1965].

1.4.2 Extension du modèle booléen

Différentes extensions ont été proposées pour remédier aux problèmes du modèle booléen standard. Dans ces extensions, chaque terme du document et de la requête est affecté par un poids. Citons :

Le modèle booléen étendu [Salton et al, 1983] : dans ce modèle, la représentation de la requête reste une expression booléenne classique. Tandis que, les termes représentant d'un document sont pondérés. L'appariement requête-document est le plus souvent déterminé par les relations introduites dans le modèle p-norme basées sur les p-distances. Où p est un facteur ajouté sur les opérateurs logiques $\vee^P$ et $\wedge^P$. La valeur de p varie dans l'intervalle [1, ∞ [. Plus p est grand, plus l'évaluation est stricte. Quand p=1, on retrouve une évaluation équivalente à celle du modèle vectoriel; Quand p→∞, l'évaluation est équivalente à celle du modèle booléen standard ou basée sur des ensembles flous.

Extension du modèle booléen basé sur les ensembles flou [Zadeh, 1965] : dans ce modèle, inspiré des ensembles flou, chaque terme a un degré d'appartenance à un document. Ce degré correspond au poids du terme dans le document. Dans ce modèle, la requête est toujours représentée par une expression booléenne classique, tandis que, l'évaluation des opérateurs logiques ∨ et ∧ et remplacer par les fonctions *min* et *max*.

Ces évaluations ont été proposées à la fin des années 1970 et au début des années 1980. Maintenant, ces extensions sont devenues standard: la plupart des systèmes booléens utilisent un de ces modèles étendus.

1.4.3 Modèle vectoriel

Dans le modèle vectoriel [Salton, 1971a], les documents ainsi que la requête sont représentés par des vecteurs dans un espace de n dimensions, constitué par les termes du vocabulaire d'indexation. Dans ce modèle, une pondération est attribuée à chaque terme de l'espace d'indexation. Ainsi, supposons E = $\{t_1, \ldots, t_n\}$ constitue l'espace d'indexation, où les t_i sont les descripteurs. L'index d'un document d_j est représenté par le vecteur de poids des termes :

$$\overrightarrow{d_j} = (w_{1j}, w_{ij}, \ldots w_{nj})$$

Où w_{ij} est le poids du terme t_j dans le document d_j. Notez qu'un terme t_k de l'espace d'indexation qui n'est pas présent dans le document, son poids sera nul ($w_{ik} = 0$). Le poids d'un terme est calculé en utilisant les méthodes de pondération citées dans la section 1.3.2 Pondération de termes.

Une requête est également représentée par un vecteur des poids $\vec{q}$ dans l'espace d'indexation E :

$$\vec{q} = (w_{1q}, w_{iq}, \ldots w_{nq})$$

Où w_{iq} est le poids du terme t_i dans la requête.

Le mécanisme de recherche consiste à retrouver les vecteurs documents qui sont les plus proches du vecteur requête. Il existe plusieurs fonctions permettant de mesurer la similarité entre deux vecteurs. Par exemple la fonction de cosinus [Salton, 1971b] est représenté par :

$$Sim_2(d,q) = \frac{\sum_i W_{di} \times W_{qi}}{\frac{1}{2}(\sum_i({W_{di}}^2) + (\sum_i {W_{qi}}^2))}$$

Le modèle vectoriel est l'un des modèles de RI classique les plus étudiés et les mieux utilisés. A l'inverse du modèle booléen qui ne permet pas de distinguer entre deux documents qui sont indexés par les mêmes termes. La fonction de pondération permet de présenter à l'utilisateur une liste des documents triés selon leur ordre de pertinence.

Plusieurs extensions ont été proposées. On peut distinguer en particulier celle basée sur une analyse sémantique latente, modèle LSI (Latent semantic Indexing) [Deerwester et al, 1990] ou encore le modèle vectoriel généralisé proposé par Wong [Wong et al, 1985] qui contrairement aux modèles classique, ne considère pas l'hypothèse d'indépendance des termes d'indexation, il permet de tenir compte des dépendances qui peuvent exister entre les termes. Le modèle LSI, quant à lui, permet de corriger les défauts du modèle vectoriel liés à la non-prise en compte des variations linguistiques (principalement synonymie) des mots. Ce modèle utilise les techniques de l'analyse en composante principale sur l'espace des mots afin de, le ramener à un espace de concepts. Pour ce faire, LSI exploite les corrélations (co-occurrences) entre les mots afin de regrouper ceux qui sont susceptibles de représenter un même concept dans une même classe. On obtient ainsi, une représentation conceptuelle des documents, ce qui limite l'impact de la variation dans l'utilisation des mots dans les documents.

1.4.4 Les modèles probabilistes

1.4.4.1 Binary Independent Model

Les modèles probabilistes sont fondés sur la théorie des probabilités [Robertson et al, 1976] [Salton et McGill, 1983] [Kuhn, 1960]. Le premier modèle probabiliste a été proposé par Maron et Kuhns [Maron & al, 1960]. Ce modèle utilise un modèle mathématique fondé sur la théorie de la probabilité. La similarité entre un document et une requête est mesurée par le rapport entre la probabilité qu'un document *d* donné soit pertinent pour une requête *Q*, notée $p(d, Q)$, et la probabilité qu'il soit non pertinent et $p(\bar{d}, Q)$. Ces probabilités sont estimées par les probabilités conditionnelles selon qu'un terme de la requête est présent, dans un document pertinent ou dans un document non pertinent. Cette mesure noté RSV(Q,d) ou aussi *O(d)* va pouvoir classer les documents selon leurs probabilités. Plus ce score est élevé pour un document, plus ce document doit être classé en haut, c'est ce que l'on appelle « principe de classement par probabilité », il est donné par :

$$RSV(Q,d) = O(d) = \frac{p(d,Q)}{p(\bar{d},Q)} = \sum_{i=1}^{k} \log \frac{p(1-q)}{q(1-q)}$$

Avec *p* désigne $P(t_i/R)$, la probabilité qu'un terme t_i de la requête apparait dans un document *d* appartenant à la liste *R* des documents pertinents par rapport à *Q*. *q* désigne $P(t_i/NR)$, la probabilité que t_i apparait dans la liste *NR* des documents non pertinents par rapport à *Q*. *k* est le nombre total de termes dans la requête.

Selon [Robertson et al. 1976], les documents et la requête sont représentés par des vecteurs dans l'espace d'indexation E = {t_1, …, t_n} comme dans le modèle vectoriel. Dans ces vecteurs les pondérations des index sont binaires. Autrement dit, les poids des termes prennent seulement les valeurs *0* (absent) ou *1* (présent). Exemple, l'index d'un document d_j est représenté par le vecteur de poids des termes :

$$\overrightarrow{d_j} = (w_{1j}, w_{ij}, \dots w_{nj})$$

Où w_{ij} = 1 si terme est présent dans le document et 0 sinon

Après quelques transformations, on arrive à la formule classique proposée par Robertson et Sparck Jones en 1976 :

$$RSV(Q,d) = \sum_{i=1}^{k} x_i * W_i = \sum_{i=1}^{k} x_i * \log\frac{p(1-q)}{q(1-q)} = \sum_{i=1}^{k} x_i * \log\frac{r_i/(n-r_i)}{(R_i-r_i)/(\mathrm{N}-R_i-\mathrm{n}+r_i)}$$

Avec x_i = 0 ou 1 qui représente l'absence ou la présence du terme t_i de la requête Q dans le document D, W_i représente le poids affecté à t_i dans le document d qui est estimé en fonction de p et q. p *et* q sont calculés en fonction des paramètres r_i, n, R_i et N. Où, N est égal au nombre totale des documents dans la collection, R_i est égal au nombre total des documents contenant le terme t_i, r_i égal au nombre de document pertinent contenant le terme t_i et n est égale au nombre de documents pertinents.

Pour éviter les 0, un lissage de cette fonction est proposé par Robertson-Sparck Jones qui est souvent utilisé dans des approches probabilistes en RI. Elle consiste à déterminer le poids d'un terme *t* :

$$W_i = \log\frac{(r_i+0.5)/(n-r_i+0.5)}{(R_i-r_i+0.5)/(\mathrm{N}-R_i-\mathrm{n}+r_i+0.5)}$$

Lorsque les données d'apprentissage ne sont pas disponibles alors on n'a pas d'informations de pertinence (n=r_i=0). On trouve le facteur *idf* probabiliste intégré dans le modèle vectoriel :

$$W_i = \log\frac{\mathrm{N}-R_i}{R_i}$$

Un des inconvénients de ce modèle est l'impossibilité d'estimer ses paramètres si des collections d'entraînement ne sont pas disponibles. Pour pallier cet inconvénient, Robertson a proposé le modèle 2-poisson basé notamment sur la notion de termes élites [Robertson, 1994a][Robertson, 1994b]. Le résultat de ces travaux est la formule BM25, largement utilisée dans les travaux actuels de RI.

1.4.4.2 **Modèle BM25**

Dans un modèle probabiliste, l'estimation de poids d'un document *D* pour une requête *q* dépend de la présence ou l'absence du terme *t* de la requête *q* dans le document *D*. Or, il a été montré que la fréquence des termes joue un rôle important dans la discrimination des termes dans les documents. Le modèle BM25 propose de tenir compte de ces fréquences dans l'estimation de différentes probabilités. Une solution proposée par Robertson, le modèle 2-poisson basé sur la notion de termes élites qui a donné comme résultat la formule BM25 [Robertson 1994a][Robertson, 1994b]. Il s'appuie particulièrement sur une hypothèse disant que la probabilité qu'un mot apparaisse k fois dans un document suit une loi de Poisson :

$$P(freq(t,D) = k) = \frac{e^{-\lambda}.\lambda}{k!}$$

Avec λ est la moyenne des occurrences du mot dans un document

k est la fréquence d'un terme t dans le document D.

De plus les termes ne jouent pas le même rôle dans les documents, il dépend de l'élite. La notion d'élitisme, consiste à distribuer les documents entre 2 groupes. Ceux qui traitent du thème représenté par le terme *t* (dans lesquels le terme *t* sera plus fréquent), noté ensemble élite (elite set), noté *E.* Ceux qui ne traitent pas le thème *t*, dans lesquels l'apparition de *t* est marginale. Les distributions de terme *t* dans les deux groupes sont différentes. La probabilité qu'un document *D* soit dans l'ensemble élite d'un terme *t* sachant le nombre d'occurrence *k* de ce terme dans le document est estimé en utilisant le théorème de Bayes.

Le modèle BM25 développé par [Robertson et al, 1994a], qui est une extension du modèle probabiliste. Il tente de prendre en compte les fréquences des mots dans les documents, avec *W* défini par :

$$W(t,d) = \frac{f_{t,d} * (k_1 + 1)}{k_1 * ((1-b) + b * \frac{dl}{avdl}) + f_{t,d}} \text{x} \log(\frac{N - df(t,C) + 0.5}{df(t,C) + 0.5})$$

Avec et $f_{t,d}$ est la fréquence du mot t dans le document *d*, *dl* est la taille du document (le nombre total d'occurrences de mots). $df(t,C)$ est le nombre de documents de la collection *C* contenant *t*. k_1 et *b* sont des constantes qui dépendent des collections des tests ainsi que du type des requêtes. *avdl* représente le moyen de longueur de tous les documents dans la collection *C*.

1.4.4.3 **Modèle statistique de langue**

Un modèle de langue tente de capturer, de modéliser la distribution des mots (des termes) dans une langue (un texte, un document, une collection de document). Il permet ainsi d'estimer la probabilité qu'une séquence donnée apparaisse dans une langue. La taille de la séquence de mots estimée par le modèle fournit le type du modèle 1-gram, 2-grams, n-grams. 1-gram estime les probabilités des séquences de 1 mot (2 mots dans les 2-grams et n mots dans le n-grams).

Ainsi, pour estimer la probabilité d'une séquence *s* de *n* mots ($s = m_1 \ldots m_n$), on utilise le modèle de langue n-grams, une approximation de dépendance limitée à n mots est utilisée:

$$P(s) = P(m_1 \ldots m_n) = \prod_{i=1}^{n} P(m_i | m_{i-k+1} \ldots m_{i-1})$$

L'interprétation du modèle de langue en RI se fait de manière naturelle, on considère en fait un document comme une langue, on estime le modèle de langue du document, ce qui revient à estimer les probabilités des séquences de mots possibles dans ce document. Ce sont souvent des séquences de 1 mot (modèle 1-gram). La pertinence d'un document vis-à-vis d'une requête est interprétée comme la probabilité que la requête formée par une suite de mots soit générée par le modèle du document. Formellement ceci est traduit de la manière suivante :

$$RSV(Q,d) = P(Q/d) = \prod_{m \in Q} P(m/d)$$

La probabilité $P(m/d)$ peut être estimée en se basant sur l'estimation par maximum de vraisemblance (*maximum likehood estimation*). Elle est donné par :

$$P_{ML}(\mathrm{m}|\mathrm{d}) = \frac{\mathrm{f(m)}}{\sum_{\mathrm{m_i} \in \mathrm{D}} \mathrm{f(m_i)}}$$

Avec f(m) = la fréquence d'occurrence de m dans le corpus

Pour pallier certains problèmes comme un n-gram qui n'apparaît pas dans le document a une probabilité nulle et ensuite, toute séquence qui le contient a une probabilité nulle. Des techniques de lissage ont été proposées, citons *le lissage de Laplace*, *lissage de Good-Turning, lissage Backoff* et *lissage par interpolation.*

D'autres approches, consistent à créer un modèle de langue pour la requête et calculer la probabilité qu'un document soit généré par ce modèle $\mathrm{P(D|M_Q)}$ [Hiemstra, 2001][Miller,1998]. Ou encore estimer les modèles de requête et de document puis comparer leur distribution en utilisant par exemple ML distinguée, un modèle de langue M_Q pour la requête et un modèle M_D pour le document, ensuite une mesure de comparaison entre les modèles, pour donner un score de pertinence à chaque document [Zhai, 2009].

1.5 Fonctionnalité additive à un SRI

Dans les différentes méthodes d'indexation présentées dans les sections précédentes, un document ou une requête sont représentés par des mots-clés. En règle générale, les documents retournés en réponse à une requête sont ceux comportant au moins un terme de cette requête. Par conséquent, un document n'ayant aucun terme de la requête n'a pas de chance d'être renvoyé même s'il est pertinent. Cette limite peut être corrigée par une reformulation de la requête.

La reformulation de requête a pour objectif d'élargir le champ de recherche d'une requête. La reformulation consiste à ajouter des termes significatifs et/ou ré-estimer leur poids. Cette reformulation peut être effectuée de manière directe ou indirecte (*retour de pertinence* ou *relevance feedback* en anglais).

Dans l'expansion directe, les mots qui sont ajoutés doivent être fortement reliés à la requête. Typiquement, on utilise un dictionnaire de synonyme ou un thésaurus, on parle de reformulation de requêtes basée sur les concepts [Khan, 2000] [Andreason et al, 2002] [Andreasen et al. 2003] [Voorhees, 1993][Navigli et al, 2003] [Qui et al, 1993] [Järvelin et al, 2004]. Les mots reliés avec des mots de la requête par certains types de relation (e.g. IS_A) sont choisis pour étendre la requête. D'autres méthodes tentent de trouver les mots qui sont fortement reliés aux mots de la requête. Dans ces méthodes, ils exploitent les cooccurrences des mots dans les textes, plus deux mots se co-occurrent, plus on suppose qu'ils sont fortement reliés [Hiemstra et al, 2001]. D'autres approches ajoutent les mots qui sont reliés à la requête entière, pas les mots qui sont reliés à des mots de la requête. Dans l'article de Qiu et Frei [Qiu et Frei, 1993], ils pensent qu'il vaut mieux choisir des mots qui sont reliés à la requête qu'aux mots individuels de la requête.

La reformulation indirecte ou relevance feedback consiste à prendre en compte le point de vue des utilisateurs sur les documents pertinents. Le principe [Rocchio, 1971], simple, repose sur le fait que l'utilisateur est le seul à savoir exactement ce qu'il cherche et donc est le plus à même de juger de la pertinence des informations retournées par un système de recherche. Partant de cette idée, les systèmes exploitant le retour de pertinence permettent de récupérer ces « jugements » utilisateurs et les exploitent pour améliorer les résultats de recherche. On peut distinguer deux types de feedback : feedback explicite, feedback implicite. Le feedback

est défini comme explicite uniquement quand les évaluateurs savent que les feedback envoyés sont interprétés comme jugement de pertinence. Tant que, le feedback est implicite dépend de comportement des utilisateurs, tels que noter les documents qui sont sélectionnés ou simplement vu par un utilisateur, le temps passé dans la lecture du document, ou l'action de défilement ou la recherche dans la page.

1.6 Evaluation des SRI

Dans les sections précédentes, nous avons décrit les facteurs de base d'un système de recherche d'information et les différents modèles qui ont été proposés dans le domaine de RI. Dans cette section, nous cherchons à évaluer la performance de ces systèmes. Cette étape constitue une étape importante qui permet de fournir des éléments de comparaison entre modèles ou approches.

Les premières questions discutées sur ce sujet concernent les critères d'évaluations et le cadre pour effectuer ces évaluations. Plusieurs critères ont été proposés, les plus utilisés sont ceux qui mesurent la capacité d'un système à sélectionner des documents pertinents. Quant au cadre, un travail important a été effectué dans le cadre du projet Granfield, qui a donné naissance au protocole qui porte toujours ce nom et qui est largement utilisé dans les campagnes d'évaluation actuelles.

Dans cette section, nous commençons par présenter le protocole d'évaluation ainsi que les collections de tests, ensuite nous présentons les différentes mesures d'évaluations qui sont utilisées dans les campagnes d'évaluations pour comparer les résultats des différents systèmes de recherche d'informations. Nous finissons par une présentation sur les campagnes d'évaluation des SRI.

1.6.1 Protocole d'évaluation

L'évaluation des SRI a débuté dans les années 1950 où des petites collections de documents (références bibliographiques) ont été utilisées à petite échelle. Dans les 1960, Cleverdon [Cleverdon, 1962] dans le cadre du projet « Cranfield project II » initie un nouveau paradigme d'évaluation des systèmes de type laboratoire (*laboratory-based model*), appelé paradigme de Cranfield. Le but étant la comparaison entre les langages d'indexation basés sur l'utilisation du même ensemble de documents, des mesures besoins/requêtes et les mêmes mesures basées essentiellement sur le rappel et la précision.

Le modèle Cranfield est fondé sur la construction de collections de test volumineuses (e.g. cranfield, CACM) servant de base à l'évaluation des systèmes différents et leurs impacts en pratique et la comparaison des différentes techniques.

La collection de test servant à l'évaluation orientée-laboratoire des SRI comprend un ensemble de requêtes, une collection de documents et des jugements de pertinence associant un sous-ensemble de documents, dits pertinents pour chaque requête :

Collection de documents : La collection de documents doit contenir des échantillons des types textes qui sont considérés représentatifs de la tâche considérée. Ils doivent donc disposés d'une diversité de document, genre, quantité, texte intégral ou résumé du texte, etc. pour qu'ils soient représentatifs d'une tâche réelle ou besoin en information.

Les requêtes : La requête traduit le besoin en information de l'utilisateur et elle est formulée souvent par un ensemble de mots clés ciblant les documents recherchés. L'évaluation d'un système ne doit pas reposer seulement sur une requête. Pour avoir une évaluation assez objective, un ensemble de quelques dizaines de requêtes, traitant des sujets

variés, est nécessaire. Vu que l'efficacité d'un système varie considérablement entre les requêtes, plus le nombre de requêtes utilisées dans les expérimentations est élevé, plus les conclusions tirées des expérimentations seront fiables [Buckley et al, 2000].

Jugement de pertinence : Dans le but de comparer les documents résultats fournis par le système et les documents que souhaite recevoir l'utilisateur, il faut spécifier pour chaque requête l'ensemble de réponses idéales du point de vue de l'utilisateur. La spécification des jugements de pertinence des documents associés à la requête constituent la tâche la plus difficile dans la construction d'une collection de test. La création des jugements de pertinence dans l'évaluation orientée laboratoire est souvent basée sur une technique appelée, *Pooling*. La technique de Pooling consiste à trouver pour chaque requête, un pool de documents constitué à partir des 100 premiers documents restitués par chacun des systèmes participant à la campagne d'évaluation, les doublons sont supprimés (opération d'union ensembliste). L'hypothèse est que le nombre et la diversité des SRI contribuant au pool permettront de trouver un maximum de documents pertinents. Dans cette technique, les jugements de pertinence sont souvent spécifiés par un groupe de personnes (assesseurs) experts dans le domaine des sujets traités par les requêtes. Dans le but d'établir les listes de documents pertinents pour chaque requête, les utilisateurs (ou des assesseurs simulant des utilisateurs) doivent examiner chaque document du pool de la collection, et juger s'il est pertinent indépendamment des autres documents pertinents contenant la même information.

1.6.2 Les mesures d'évaluations

1.6.2.1 Les mesures de rappel précision

Pour mesurer la pertinence d'un SRI en terme d'efficacité c'est-à-dire sa capacité à trouver des documents pertinents, 2 mesures ont été définis : le *rappel* et la *précision*.

Le *rappel* est défini par le nombre de documents pertinents retrouvés sur le nombre de documents pertinents de la requête. La *précision* est le nombre de documents pertinents retrouvés rapporter au nombre de documents total proposés par le moteur de recherche pour une requête donnée. Un système est dit *précis* si peu de documents inutiles sont proposés par le système, ce qui signifie que le taux de *précision* est élevé. Le taux de *rappel* et de *précision* sont mesurés par les formules suivantes :

$$rappel = \frac{nombre\ de\ documents\ pertinents\ retrouvés}{nombre\ de\ documents\ pertinents\ de\ la\ collection}$$

$$précision = \frac{nombre\ de\ documents\ pertinents\ retrouvés}{nombre\ total\ de\ documents\ retrouvés}$$

Un système qui aura à la fois le rappel et la précision égal à 1, cela signifie qu'il a trouvé tous les documents pertinents et rien que les documents pertinents. Cette situation est très loin de la réalité. Le plus souvent, il est possible d'obtenir un taux de rappel et de précision aux alentours de 30%.

1.6.2.2 La courbe précision-rappel

Les mesures de précision-rappel ne sont pas indépendantes, en effet en réponse à une requête on a un taux de rappel égale à 1, mais une précision faible, voir de même, si on augmente la précision en restreignant le nombre de documents retournés, dans ce cas le rappel pouvant diminuer. Dans les SRI on cherche à améliorer le couple rappel et précision.

Ces deux métriques ne sont pas statiques non plus (c'est-à-dire qu'un système n'a pas qu'une mesure de précision et de rappel). Le comportement d'un système varie en fonction de

précision et de rappel donc de la liste ou du rang du document dans la liste. Ainsi, la courbe de la Figure 4 montre la forme générale que peut prendre la variation de rappel précision pour un système.

La précision et le rappel croient en même temps si un document pertinent est récupéré. Plus cette courbe décroît tardivement, meilleur est l'algorithme de l'ordonnancement étudié. Cette courbe est intéressante pour comparer les résultats d'une même requête rendus par deux SRI différents. Un système dont la courbe dépasse (c'est-à-dire qu'elle se situe en haut à droite de) celle d'un autre est considéré comme un meilleur système. Il arrive parfois que les deux courbes se croisent. Dans ce cas, il est difficile de dire quel système est le meilleur.

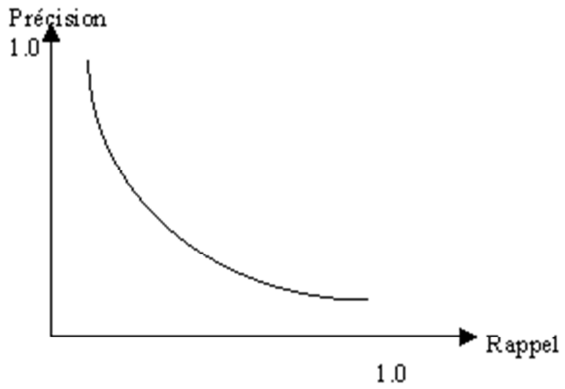


Figure 4: Courbe de précision-rappel

1.6.3 La R-précision et la MAP (Mean Average Precision)

La R-précision ou la précision exacte représente la précision calculée pour le R^{ème} document pertinent retourné si la requête admet R documents pertinents dans la base. La précision à x documents est la précision calculée au x^{ème} document retourné (x= 5, 10, 10, etc.).

Comparons à la courbe de précision, la MAP présente l'avantage qu'elle permet de décrire la performance globale d'un système. C'est la mesure, la plus souvent utilisée en RI, ce qui est le cas notamment dans les cadres de TREC et de CLEF. Elle tient compte à la fois de la précision et du rappel. Elle est formée par une moyenne des AP (Average Precision) calculée pour chaque requête testée par le système, avec AP représente la moyenne des précisions (non interpolées) calculées pour chaque document pertinent à trouver, au rang de ce document. Si un document pertinent est retourné à la dixième position, la précision pour ce document est « la précision à 10 documents ». Si un document pertinent n'a pas été trouvé par le système, la précision pour ce document est nulle. Pour une requête donnée, la précision moyenne, noté AP est calculée comme suit :

$$AP = \frac{1}{R} \sum_{i=1}^{N} p(i) \times R(i)$$

Où R(i) = 1 si le i^{ème} document restitué est pertinent, R(i) = 0 si le i^{ème} document restitué est non pertinent, *p(i)* la précision à *i* documents restitués. *R* le nombre de documents pertinents restitués et *N* le nombre total de documents retournés par le système.

Ainsi MAP est calculée pour un ensemble *C* de requête traité par le système comme suit :

$$MAP = \frac{1}{k} \sum_{q \in C} AP$$

Avec k = |C|, le nombre de requête dans C.

1.6.4 Présentation des campagnes d'évaluation des systèmes de recherche d'informations

Pour tester l'efficacité et la performance de systèmes de recherche d'information, des campagnes d'évaluation ont été mises en place depuis les années soixante à Cranfield (Royaume-Uni). Suivi quelques années plus tard par le projet MEDLARS (Medical Literature Analysis and Retrieval System), réalisé à la bibliothèque nationale de médecine aux Etats-Unis. En 1992, les campagnes TREC (Text REtrieval Conference) créées par le NIST (National Institut of Science and Technology) sont devenues la référence en ce qui concerne l'évaluation des systèmes. On peut citer les campagnes CLEF (Cross-Language Evaluation Forum) qui se rattachent plus particulièrement aux systèmes multilingues, les campagnes NTCIR sur les langues asiatiques et Amaryllis, spécialisées sur le système français.

1.6.4.1 La campagne d'évaluation TREC

Suite aux expérimentations du programme Cranfield [Cleverdon, 1962], la campagne d'évaluation TREC[2] est née depuis 1992. TREC c'est une série d'évaluation annuelle des technologies pour la recherche d'informations crée par le NIST dans le but de proposer des moyens homogènes d'évaluation de systèmes documentaires et co-sponsorisés par le NIST et la DARPA (Defense Advanced Research Projects Agency) sur des bases de documents conséquentes. Aujourd'hui, les campagnes TREC sont devenues la référence dans l'évaluation.

Les objectifs de ces campagnes étaient de favoriser l'application des techniques de RI sur des grandes collections de données, de favoriser l'échange de technologie entre les industriels et la communauté scientifique, de comparer les performances des différentes techniques, et enfin de définir un protocole d'évaluation homogène pour toute la communauté de RI.

Les participants à ces campagnes cherchent à améliorer la performance de leurs systèmes. Les conditions de participation sont les suivantes : le NIST diffuse courant décembre un appel à participation qui explique dans les grandes lignes les objectifs et le déroulement des tâches (testes) pour l'année à venir. Les demandes de participation doivent être déposées en Janvier. La participation à la conférence annuelle elle-même est soumise à l'envoi au NIST des résultats.

La campagne TREC offre d'importantes collections de tests et un ensemble de tâches diverses qui évoluent d'une année à l'autre. Les expérimentations qui y sont menées sont complètes. La première campagne TREC (TREC-1) a vu le jour en 1992 avec 25 participants issus du monde académique et industriel. Un certain nombre de tâches existent dans TREC et sont dédiées à certains aspects de la recherche tels que la génomique, le web, le filtrage, les blogs, les questions réponses, etc. L'édition 2011 de TREC comprend plusieurs tâches dont la tâche blog track qui s'intéresse à l'étude des informations contenues dans la "blogosphère", la tâche Web track correspond aux tâches de récupération spécifiques au Web, y compris les tâches de la diversité et l'efficacité, plus de collection allant jusqu'à un milliard de pages, etc.

À chaque session, TREC met à disposition des participants à la campagne un ensemble de documents et de requêtes. Pour chacune des requêtes, une liste des documents pertinents est déterminée par des juges humains. TREC met aussi à disposition des participants un programme nommé trec-eval qui permet de calculer, pour un ensemble de requêtes, les performances des systèmes selon plusieurs critères et mesures. Plusieurs tâches sont traitées

2 http ://trec.nist.gov

dans les campagnes TREC, citons les principales : filtrage, recherche (ou tâche ad hoc), interrogation interlingue et question-réponse.

La principale tâche est la tâche adhoc. Dans cette tâche, l'évaluation se fait en comparant les documents pertinents restitués par les divers systèmes participants pour une requête, et la liste des documents pertinents jugés par les juges de TREC pour la même requête testée, en utilisant le programme trec-eval. Le programme trec-eval calcule pour les 1000 premiers documents, les performances des listes restituées par les systèmes.

Le corpus de documents utilisé par la tâche ad-hoc s'agit d'une collection fermée des documents dont le fonds documentaire s'enrichit chaque année depuis 1992. Il est composé en majeure partie d'articles de presse, concernant l'actualité dans tous les domaines : périodiques, quotidiens, dépêches de presse (Financial Times, San Jose Mercury News, Wall Street Journal, Ziff-Davies publications, Associated Press Newswire, Foreign Broadcast Information Service, Los Angeles Times, etc.) et de textes juridiques ou documents officiels : brevets du US Patents and Trademark Office, textes réglementaires issus du Federal Register, littérature du Département of Energy, procès-verbaux des séances du Congrès, Congressional Report (Federal Register). Le corpus du TREC comporte aujourd'hui, plus de 5 millions de documents textuels. Lorsque la requête est courte (2 ou 3 mots), la tâche ad hoc ressemble à la recherche d'informations sur le web.

Dans les évaluations, une collection de 500,000 à 700,000 documents est fournie pour chaque utilisateur par le NIST qui doit l'indexer par sa propre méthode d'indexation. Avec ces documents, le NIST procure également aux participants un ensemble de 50 requêtes. Pour chaque requête, les participants classent les documents par ordre de pertinence. Les 1000 premiers documents retirés par les différents systèmes pour chaque requête sont soumis à NIST. Les examinateurs font une évaluation de pertinence des 100 ou 200 premiers documents de chaque système et attribuent à chacun différents scores d'évaluation.

1.6.4.2 **Autres campagnes d'évaluations**

D'autres campagnes d'évaluation existent :

CLEF [3]: la campagne d'évaluation CLEF est lancée en 2000, dans le cadre d'un projet européen, pour l'évaluation des systèmes de recherches d'informations qu'ils soient monolingues ou multilingues de langues européennes. Depuis 2002, il intègre la campagne d'évaluation des systèmes de recherches textuelles pour la langue française « Amaryllis ».

Les tâches principale proposées par CLEF sont les tâches monolingue, bilingue, multilingue et les recherches dans un domaine spécifique. En CLEF 2004, six tâches sont proposées : recherche mono-, bi-, multilingue dans de nouvelles collections (ad-hoc), recherche mono-, bi-, multilingue dans un domaine spécifique (GIRT), recherche inter lingue interactive (iCLEF), question/réponse multilingue (QA@CLEF), recherche inter lingue dans une collection d'images (Image CLEF) et recherche inter lingue dans des documents audio (CL-SDR).

NTCIR[4] : en 1999, les campagnes NTCIR (NII-NACSIS Test Collection for IR Systems) sont apparues dans le but d'améliorer tous les domaines de l'accès à l'information y compris la recherche d'information, la production de résumés, l'extraction terminologique, etc. La collection de test utilisée comprend des textes publiés en 1998 et 1999, en chinois traditionnel, en coréen, en japonais et en anglais.

3 http ://www.clef-campaign.org/

4 http ://research.nii.ac.jp/ntcir/

1.6.4.3 Limites des campagnes d'évaluation des systèmes de recherches d'informations

Malgré que, les compagnes d'évaluations ont amélioré l'efficacité des systèmes, on peut citer certaines limites. Soit au niveau de corpus de documents qui est seulement thématique et non logique et ne prend pas en compte l'opinion de l'auteur. Soit au niveau de l'évaluation faite uniquement par des professionnels. Soit au niveau de l'évaluation qui est faite par rapport au nombre des documents retrouvés, or, en général, un utilisateur ne cherche pas des documents, mais de l'information et la quantité d'information dépend d'un document à un autre. D'autres limites sont par rapport au corpus de requêtes, où les requêtes qui doivent représenter un véritable besoin d'information tel qu'il est perçu par l'usager sont composées par des mots clés. C'est un problème général en RI, il est relié à la représentation du besoin d'information de l'usager et de savoir poser les questions à ces systèmes. Une autre limite est reliée aux jugements de pertinence qui est une notion subjective, un document dépend de la personne qui le juge. Une limite importante, un document est jugé pertinent ou non pertinent, pourtant, certains documents sont plus pertinents que d'autres suivant le besoin de l'usager.

1.6.5 Les outils d'évaluations

Plusieurs outils d'évaluations ont été créés pour la recherche d'information, qui utilisent différents modèles pour indexer une collection, Michel Beigbeder dans son site[5], représente la plupart de ces outils. Citons par exemple SMART, qui implémente le modèle vectoriel basique avec la possibilité d'expérimenter les différents schémas de pondération. SMART est écrit en *langage C* qui est en développement continu depuis 1992. Mercure [Boughanem et al, 2003], est un autre système développé au sein de notre équipe, qui, lui aussi implémente le modèle vectoriel. Nous utilisons Mercure dans nos expérimentations pour comparer les résultats d'OKAPI avec nos résultats. D'autres logiciels, tels que, TERRIER implémenté en Java pour le développement rapide de Web et pour les moteurs de recherches intranet et bureaux. TERRIER propose des stratégies d'indexation multiples, telles que l'indexation MapReduce multi-pass, un seul passage et à grande échelle.

1.7 Conclusion

La recherche d'information est l'ensemble des méthodes, procédures et techniques permettant d'identifier à partir d'un ensemble de documents les informations pertinentes susceptibles de répondre à la requête. Dans ce chapitre, nous avons montré le concept de la recherche d'information et le fonctionnement d'un système de recherche d'informations. Nous avons présenté également, ait le processus de représentation ou d'indexation de l'information et celui de l'appariement document- requête ait les différents modèles utilisés dans le domaine, ainsi que, les campagnes d'évaluations pour tester la capacité d'un SRI et améliorer ses performances.

Dans ce chapitre, on a décrit le fonctionnement d'un SRI général. Dans le second chapitre, nous focalisons sur les méthodes de recherche d'informations conceptuelles et sémantiques qui sont en relation avec notre proposition.

5 http://www.emse.fr/~mbeig/IR/tools.html

Chapitre 2 Représentation des textes

2.1 Introduction

Un texte peut être représenté par deux différentes techniques. La première consiste à le représenter par un sac de descripteurs (mots simple, multi-terme, concepts,…) correspondant à des mots qui le composent, ce processus nommé indexation permet de trouver les éléments significatifs, souvent des mots clés, qui représentent le contenu d'un document. C'est une des opérations importantes qu'on trouve dans la majorité des tâches de RI, tels que la recherche adhoc, la classification de documents, le résumé des documents, etc. Les descripteurs ainsi extraits peuvent avoir différentes formes (mots clés, mot-sens, concepts) et peuvent être représentés sous forme d'un ensemble de mots ou encore sous forme de structure plus complexe (graphe, hiérarchie, …).

La deuxième consiste à le résumer, où on préserve les informations importantes du contenu original. Les méthodes de production de résumés automatiques de texte peuvent être groupées en deux familles : extraction et abstraction. Les systèmes produisant des résumés par abstraction sont fondés sur la compréhension du document et la génération d'un véritable texte grammatical et cohérent. L'approche par extraction consiste en la sélection des unités (mots, phrases, paragraphes, etc.) censées contenir l'essentiel de l'informativité du document et en la production d'un extrait par assemblage de ces dernières.

Dans ce chapitre, nous présentons un état de l'art sur les différentes techniques de représentation d'un texte. Nous commençons par une description des différentes techniques d'extraction des descripteurs. Ensuite, nous présentons WordNet, une ressource terminologique qui permet d'identifier les relations sémantiques existantes entre les termes d'un même document. Nous décrivons les différentes méthodes de représentation des descripteurs par cluster, réseaux, graphe, etc. Nous parlons ensuite des différentes méthodes de résumée textuelle, précisément sur les résumées par extraction en relation avec nos travaux. Nous terminons ce chapitre par une conclusion générale.

2.2 Différentes formes qui peuvent avoir un descripteur

Nous présentons dans cette section les différentes formes qui peuvent avoir un descripteur, ainsi que les différentes approches utilisées pour l'extraction de ces descripteurs.

2.2.1 Unité représentative d'un descripteur

Les descripteurs constituent l'information atomique de représentation d'un texte [Salton et McGill, 1983]. Ils peuvent prendre les formes suivantes:

- *Mots simples* : les textes sont simplement transformés en un vecteur de leurs mots simples en éliminant les mots vides ("de", "un", "les", etc. pour la langue française, "of", "the", etc. pour la langue anglaise).
- *Racines lexicales* : les textes sont uniquement représentés par les racines des mots (on parle de *stem* en anglais). Les algorithmes proposés consistent à substituer les mots par leurs racines. L'un des algorithmes le plus connu pour la langue anglaise est l'algorithme de Porter[6] [Porter, 1980].

6 Plusieurs implémentations rapide et efficace de cet algorithme sont possibles sur le web : http://www.muscat.com/~martin/stem.html

- *Lemmes* : les mots extraits du texte sont remplacés par leurs formes d'origine. Un processus de lemmatisation consiste à faire une analyse grammaticale afin de remplacer les verbes par leurs formes infinitives et les noms par leurs formes singulières. La lemmatisation est donc plus compliquée à mettre en œuvre que la recherche de racines, puisqu'elle nécessite une analyse grammaticale des textes. Un algorithme efficace nommé *TreeTagger*[7] [Schmidt, 1994] pour les langues anglaise, française, allemande et italienne a été développé. Dans cet algorithme, l'analyse grammaticale est effectuée en utilisant des arbres de décision puis des fichiers de paramètres spécifiques à chaque langue.
- *N-grammes* : selon [Baziz, 2005], les n-grammes sont une représentation originale d'un texte en séquence de n caractères ou mots consécutifs. On trouve des utilisations de bigrammes et trigrammes dans la recherche documentaire (ils permettent de reconnaître des mots d'une manière approximative, ainsi, de corriger des flexions de mots ou même des fautes de frappe ou d'orthographe). Ils sont aussi fréquemment utilisés dans la reconnaissance de la langue d'un texte [Harbeck et al, 1999], [Dunning, 1994].
- Les *groupes de mots* : un groupe de mots est souvent une suite de mots. Une suite de mots est plus riche sémantiquement pour représenter le contenu du document que des mots prises séparément. Prenant par exemple l'expression "recherche d'information", elle est plus précise que "recherche" et "information" pris séparément.
- *Concepts* : on utilise les représentations conceptuelles comme des intermédiaires entre le langage libre utilisé par un utilisateur et un langage contrôlé qui est souvent une liste de concepts (on parle de vocabulaire contrôlé). Cette liste peut être décrite dans un thésaurus, une ontologie, une hiérarchie de concepts, etc. Les concepts sont des unités élémentaires de cette liste et peuvent prendre la forme des expressions contenant un ou plusieurs mots.
- Les *contextes* : dans ce cas les descripteurs sont des termes qui n'apparaissent pas forcément dans le texte du document mais ayant un lien sémantique et/ou de cooccurrence avec les mots du document. Les concepts qui sont ajoutés ayant un score de similarité sémantique avec les concepts qui décrivent le document [Khan, 2000]. La disponibilité des ontologies et des ressources externes facilite l'extraction de ces concepts. L'analyse de cooccurrences des mots dans un corpus est utilisée dans le cas de "Latent Semantic Indexing" [Deerwester et al, 1990][Baziz, 2005]. Les documents et leurs mots sont alors représentés sur d'autres dimensions où les mots apparaissant dans un même contexte sont supposés proches sémantiquement.

Différentes approches sont proposées pour l'extraction des descripteurs représentatifs d'un texte [Claveau, 2003]: les approches numériques ou statistiques, les approches linguistiques ou structurelles, les approches hybrides ou mixtes et les approches sémantiques. Les approches numériques exploitent la nature fréquentielle des descripteurs à acquérir et utilisent le plus souvent des techniques statistiques. Les approches linguistiques utilisent des informations structurelles et symboliques pour exploiter les connaissances linguistiques. Les approches combinant les 2 premières approches sont appelées approches hybrides ou mixtes. Tant que les approches sémantiques consistent à détecter le sens des mots suivant le contexte de leurs apparitions.

7 Les codes sources relatives à cet algorithme ainsi que les publications sont disponibles sur le site : http://www.ims.uni-stuttgart.de/projekte/corplex/DecisionTreeTagger.html

2.2.2 Approches statistiques

Les approches statistiques (aussi nommés numériques) d'extraction de termes sont employées avec un grand succès dans l'indexation depuis de nombreuses années. Les techniques utilisées sont devenues plus robustes avec l'accessibilité des documents informatisés et la facilité de construction des corpus volumineux. Ces approches se basent sur des formules statistiques et sur de simples calculs de fréquence. Leurs avantages, elles sont simples à mettre en œuvre, ne nécessitent pas une connaissance spécifique des langues des corpus, ni des domaines couverts par ces corpus.

Les termes peuvent être simples ou termes complexes. Les termes simples sont les unités textuelles simples composant un document, comme un nom par exemple. La technique générale d'identification des termes simples consiste à identifier les termes qui constituent le document. Quant aux termes complexes, ils sont constitués d'au moins deux unités lexicales simples. Les techniques de détection des termes complexes nécessitent une analyse syntaxique plus fine. Nous présentons un aperçu général sur deux familles de ces techniques à base fréquentielles : les approches par cooccurrences et les approches par segments répétés.

Approches par cooccurrences : les travaux menés dans ce domaine cherchent à associer les mots qui apparaissent ensembles dans un texte de manière statistiquement significative. Les techniques utilisées se basent sur la probabilité que deux mots apparaissent ensembles dans une certaine fenêtre de texte [Church & Hanks, 1989][Church & Hanks, 1990][Church et al, 1991]. Le principe de fonctionnement de ces techniques consiste à mesurer pour chaque couple de mots un indice statistique "score" pour mesurer la force du lien unissant ces deux mots. Le couple de mots retenu est le couple qui a un score dépassant le seuil fixé.

Approches par segments répétés : [Lebart & Salem, 1994] proposent une technique dite des segments répétés. Son fonctionnement consiste à identifier dans un texte tout segment (une suite des unités textuelles) qui apparaît fréquemment dans un corpus. Dans ses travaux de thèse, Oueslati [Oueslati, 1999] utilise le principe de segments répétés pour la réalisation d'un système d'aide à la construction de terminologie d'un domaine spécialisé, telle que la médecine. Dans une première étape, il fait appel à la méthode basée sur les segments répétés pour extraire les termes. Dans une seconde étape, les termes extraits sont validés par un linguiste ou terminologue. Ensuite, il cherche à construire des classes de termes sémantiquement proches.

2.2.3 Approche syntaxique

L'extraction des termes simples et des termes composés nécessite une connaissance parfaite des règles syntaxiques de dérivation dans la langue du corpus. Les techniques utilisées dans ces approches exploitent des sources linguistiques pour obtenir une définition des structures porteuses des termes. Dans ces techniques des définitions opérationnelles des termes, ou d'objets linguistiques proches, sont établies par des linguistes puis utilisées pour trouver les candidats-termes répondant à ces définitions [Claveau, 2003][Bourigault, 1996] [Aussenac-Gilles et al, 2000]. Le plus souvent, un ensemble de patrons syntaxiques comme (NOM NOM) ou (NOM PREP NOM) est utilisé pour l'identification. Exemples des outils qui sont employés dans l'application de ces approches :

TERMINO [David et Plante, 1991] qui compte parmi les premiers outils opérationnels d'extraction des termes. Il est construit sur la base d'un formalisme pour l'expression de grammaire de la langue naturelle. La version actuelle de TERMINO se nomme NOMINO [Perron, 1996].

LEXTER de Bourigault [Bourigault, 1994][Harrathi, 2009] est un outil dédié initialement à l'enrichissement des thésaurus à partir d'un système d'indexation automatique des corpus textuels, par la suite, LEXTER a été utilisé pour l'extraction et la modélisation des connaissances à partir de corpus textuels en langue française.

FASTER [Jacquemin, 1997] est un outil dédié à connaître les termes qui apparaissent dans un corpus. Les termes identifiés sont choisis à partir d'une liste de termes fournis au système en se basant sur des analyses syntaxiques.

TreeTagger [Schmid, 1994], est un outil qui permet d'analyser une phrase ou un texte syntaxiquement afin de remplacer les verbes par leurs formes infinitives et les noms par leurs formes singulières. Il attribue donc à chaque mot sa catégorie grammaticale. L'algorithme développé est efficace pour les langues anglaise, française, allemande et italienne. Dans cet algorithme, l'analyse grammaticale est effectuée en utilisant des arbres de décisions puis des fichiers de paramètres spécifiques à chaque langue. Par la suite, dans notre proposition nous utilisons *TreeTagger* pour détecter les noms dans un texte.

Il existe d'autres outils comme KEA de Jones et Paynter, AZ NOUN PHRASER de l'université de l'Arizona, TERMS de J. Justeson et S. Katz [Justeson & Katz, 1995] etc.

2.2.4 Approches hybrides ou mixtes

Ces approches combinent les deux approches présentées précédemment (les approches statistiques et les approches syntaxiques). L'ordre de cette combinaison varie d'un système à un autre. En effet, dans certains systèmes, les résultats obtenus par une analyse linguistique sont validés et filtrés par une analyse statistique, tandis que, dans d'autres systèmes, les résultats de l'analyse statistique sont validés par une analyse linguistique [Harrathi, 2009]. Exemples des outils qui sont employés dans l'application de ces approches :

ACABIT [Daille, 1994] est un outil dédié uniquement à l'extraction des termes composés à partir du corpus. Daille, dans son algorithme reprend les techniques syntaxiques empruntées par TERMINO et LEXTER. Ensuite, il utilise les techniques statiques afin de déterminer le degré de liaison entre les mots associés dans les termes composés extraits précédemment.

XTRACT [Smadja, 1993] est un outil crée dans un travail sur l'indexation automatique. Il consiste à repérer des structures prédéfinis à partir des collections telles que : nom+nom, nom de nom, nom+adjectif, sujet+verbe, etc. Dans un premier temps, XTRACT exploite les techniques statistiques essentiellement basées sur l'information mutuelle entre les mots et dans un deuxième temps, il utilise des techniques linguistiques [Harrathi, 2009].

2.2.5 Approches sémantiques

Ces approches consistent à détecter le sens d'un terme suivant son contexte d'apparition. Dans ces approches, l'extraction des termes des documents s'appuie sur des algorithmes de désambiguïsation de mots (WSD) [Sanderson, 1997][Krovetz, 1997][Michalcea et al, 2004][Baziz, 2005]. Les termes extraits d'un document sont représentés par des concepts représentant leurs sens plutôt qu'avec des termes simples. Ces approches sémantiques prennent en considération l'ordre d'apparition des termes ce qui n'était pas le cas dans les méthodes d'extraction classiques.

Plusieurs travaux se sont intéressés à l'aspect sémantique des termes extraits. Pour désambiguïser le sens, certaines approches utilisent des représentations hiérarchiques pour calculer la distance sémantique ou similarité sémantique entre les termes à comparer [Resnik, 1995] [Leacock et al, 1998] [Hirst et al, 1998] [Jiang et al, 1997] [Lin, 1998]. Yarowsky

[Yarowsky, 1995] associe aux termes extraits, des termes du contexte qui aident à déterminer leurs sens. D'autres approches se sont basées sur les synsets de WordNet et ont amené à une amélioration remarquable dans le rappel et la précision [Mihalcea et al, 2000][Gonzalo et al, 1998].

2.2.6 Approches conceptuelles

Les approches conceptuelles consistent à rattacher les termes à des concepts sous-jacents. Dans ces approches, il faut une "liste" de concepts cibles (qui exprime le sens des termes possibles) pour pouvoir transformer le terme en concept. Les concepts sont tirés d'une liste prédéfinie des termes d'index ou un vocabulaire contrôlé appelé système de classification construite dans le but de gérer une collection de documents, qui en général intègre une structure sémantique. Des exemples communs des systèmes des classifications, les dictionnaires de synonymes, les ontologies, les thésaurus, les taxonomies, etc. [Woods, 1997][Stairmand et al, 1996][Mauldin, 1991][Stein et al, 1997].

L'avantage de représenter un texte par des termes d'un langage contrôlé, c'est que les termes sont de nature non ambigus et ont une définition précise. Le fait que les termes produisent un sens non ambigu, ils peuvent être traduits dans d'autres langages pour extraire des textes de différentes langues. En plus, ces termes représentent un point d'accès général pour des classes de textes, ce qui facilite leurs emplois dans les recherches génériques, routages et filtrages des documents par rapport à des classes générales. Cependant, la construction manuelle et l'entretien des sources des connaissances (par exemple, thésaurus) sont des tâches couteuses, et la construction automatique des thésaurus reste une tâche très difficile. En plus, les vocabulaires de langage contrôlé doivent être régulièrement mis à jour pour tenir en compte des changements dans l'intérêt et des concepts de recherche, et des changements dans la collection des documents. Cependant, même avec la mise à jour régulière, ces vocabulaires sont lents à s'adapter à l'évolution des besoins des utilisateurs dans un monde où les textes sont produits à un rythme croissant sans cesse.

Les différentes techniques utilisées correspondent à attacher les termes d'un texte aux concepts d'une ontologie. Khan [khan, 2000] se base sur la notion de région d'ontologie et de distance sémantique entre concepts pour attacher les termes aux concepts. Baziz [Baziz et al, 2004] attache les termes d'un texte aux concepts de WordNet en se basant sur la notion de la similarité sémantique entre concepts.

Ces approches appliquées à la représentation des documents, représentent les documents par un ensemble des concepts et relations extrait à partir d'une ontologie [Boubekeur et al, 2008]. Comparées aux techniques classiques qui représentent le document par un sac de mots, l'exploitation des relations entre les termes permet à un SRI par exemple, de trouver non seulement les documents qui contiennent les termes de la requête mais également, des documents contenant les mots qui sont en relation avec les mots de la requête. Les relations entre termes étant extraites d'ontologies [Khan, 2000], [Baziz et al, 2004] ou découvertes dans les contextes des documents, au moyen des règles d'association [Haddad, 2002].

Plusieurs travaux ont été menés sur les approches conceptuelles et ils sont appliqués dans plusieurs domaines. Les systèmes Textpresso [Müller et al, 2004] et MetaMap (UMLS) sont employés dans le domaine médical. De même, autres domaines utilisent l'appariement des concepts, tel que le domaine de sport [Khan, 2000], le domaine légale [Stein, 1997] ou encore dans des domaines plus généraux, comme le cas du système FERRET [Mauldin, 1991][Woods, 1997][Aggarwal et al, 2001].

2.2.6.1 **Méthode de désambiguïsation proposée par Baziz**

Baziz [Baziz, 2005] a également proposé une méthode de désambiguïsation intéressante nous la détaillons dans ce qui suit car nous l'utilisons dans nos travaux. Dans cette méthode, on identifie pour chaque terme extrait d'un texte son sens en se basant sur l'ontologie de WordNet. L'approche comprend deux étapes principales. La première consiste à identifier les termes simples et composés attachés aux documents et qui correspondent à une entrée dans WordNet. Dans la deuxième étape, on se sert de WordNet pour récupérer les différents sens possibles des termes retenus dans le but de choisir pour chaque terme un sens unique qui lui correspond dans le texte.

Avant de rentrer dans les détails de cette section, nous précisons tout d'abord la notion de *concept* utilisé dans le cadre de ces travaux. Un concept est en fait, une entrée de l'ontologie WordNet notée « synset » comme décrite dans la section 2.3.1. Ainsi que, nous utilisons la notion *concept-terme* pour désigner un sens d'un terme appartenant à un concept WordNet.

Plus en détail, dans la première étape l'objectif est d'extraire tous les termes du document susceptibles de représenter des concepts de WordNet. Pour cela, avant d'éliminer tous les mots vides du document (ceux de la *stoplist*, "of", "the", etc.), la méthode tente de détecter les termes composés par des mots uniques ou des groupes de mots. Ces termes peuvent correspondre à différentes entrées (ou nœuds) dans l'ontologie. Les termes ainsi identifiés, peuvent être des groupes nominaux comme *pull_one's_weight* ou des entités nommées telles que *united_kingdom_of_great_britain_and_northern_ireland*. Plus précisément Baziz propose de concaténer des mots adjacents dans le texte et voir s'ils correspondent à une (des) entrée(s) dans l'ontologie, s'ils ne correspondent pas à des entrées dans l'ontologie, on utilise leurs formes de base.

Prenons par exemple cette phrase: "*The **domestic dog** has been one of the most widely kept working, hunting and companion animals in human history*."

Dans cette phrase en combinant les mots adjacents « domestic » et « dog », ils correspondent à un concept de WordNet « domestic dog » donc nous considérons le terme « domestic_dog » comme un multi-terme du document.

En fait, la détection des multi-termes permet de réduire l'ambiguïté lors de l'affectation d'un terme à un concept de l'ontologie. En général, les multi-termes sont monosémiques (ils n'ont qu'un seul sens) même si les mots qui les composent peuvent être ambigus. Si nous prenions par exemple chaque mot séparément, dans le terme *ear_nose_and_throat_doctor*, nous aurions à désambiguïser entre 5 sens pour le mot *ear*, 13 sens (7 pour le nom et 6 pour le verbe) pour le mot *nose*, 3 sens pour *throat* (and étant un mot vide, il n'est pas utilisé) et 7 sens (4 pour le nom et 3 pour le verbe) pour le mot *doctor*. Tandis que *ear_nose_and_throat_doctor* correspond à un seul sens.

Pour clarifier cette étape, on prend un exemple sur un document *d.* Après l'extraction de ces termes simples et composés, le document sera représenté par un ensemble des termes R_d défini par :

$$R_d = \{t_i \,/\, i = 1, k(d)\}$$

Avec t_i est un terme simple ou composé (multi-terme) du document *d* qui correspond à une entrée dans WordNet. *k(d)* est le nombre des termes dans le document *d*.

Le problème par la suite, c'est de trouver pour chaque $t_j \in R_d$ un unique sens ou son *concept-terme* qui appartient à un *concept* (synset) de WordNet. Cette étape est essentielle puisqu'on utilise WordNet comme ontologie, et WordNet est très précise dans la couverture de sens des mots, ce qui implique qu'un terme peut être représenté par plusieurs sens

possibles. Cette couverture de sens complique la tâche de désambigüisation. Exemple, le verbe *« to give »* n'a pas moins de 44 sens, ce qui donne qu'il faut une tâche de désambigüisation lexicale sophistiquée pour détecter le sens qui correspond à un contexte.

La deuxième étape consiste à utiliser une méthode de désambiguïsation pour identifier pour chaque $t_j \in R_d$ le concept-terme qui lui correspond dans WordNet. Il existe différentes méthodes de désambigüisation de mots (WSD) basées sur WordNet [Sanderson, 1997][Krovetz, 1997][Michalcea et al, 2004][Baziz, 2005].

Dans cette section, nous décrivons brièvement la méthode proposée par Baziz [Baziz, 2005]. Dans un premier temps, on se sert de WordNet pour récupérer les différents sens possibles pour les termes retenus. Puis, pour chaque terme du document, on calcule un score pour chacun des sens possibles (concepts candidats). Le calcul de score se base sur les mesures similarités entre les différents sens des termes (concepts candidat) constituant un document en utilisant des mesures tels que les mesures de Leacock et Chodorow (ou Lch) [Leacock et al, 1994], de Lin [Lin, 1998], de Resnik [Resnik, 1999] et la mesure de Patwardhan, Banerjee et Pederson [Patwardhan et al, 2003]. Le score d'un concept candidat est obtenu en sommant les valeurs de similarité qu'il a avec les autres concepts candidats (correspondant aux différents sens des autres termes du document). Les concepts candidats ayant les plus grands scores sont alors sélectionnés pour représenter le document.

Poursuivant l'exemple sur le document *d*. Pour identifier les concepts qui composent *d*, on projette R_d sur l'ontologie O (WordNet dans notre cas). Pour chaque terme $t_i \in R_d$ on le représente par un concept unique de O. Notant $N(O, d)$ l'ensemble des concepts qui constituent un document :

$$N(O,d) = \{c_j \,/\, j = 1, m(d)\}$$

Avec, pour chaque concept $c_j \in N(O, d)$, $\exists\ t_j \in R_d$ qui lui est associé, et c_j correspond à un *concept-terme* de WordNet. $m(d)$ est le nombre de concepts de O qui sont équivalents à des termes dans R_d.

2.3 WordNet, ressources externes pour l'extraction des descripteurs

Depuis les années 1990, avec l'apparition de grandes bases de connaissances mises en place pour répondre aux besoins en ressources sémantiques lexicales élaborées, le domaine de la construction de lexiques sémantiques en Traitement Automatique des Langures « TAL » a connu un essor. Ces bases ont commencé à grandir sans cesse pour couvrir un maximum de domaines. Ces ressources sont utilisées dans l'indexation sémantique ou conceptuelle décrite précédemment. La qualité de l'indexation dépend de la richesse sémantique des ressources utilisées. Dans cette section nous présentons WordNet, une des ressources sur laquelle nous nous sommes appuyées dans le cadre de nos travaux.

WordNet est une base de données lexicales développée par des linguistes du laboratoire des sciences cognitives de l'université de Princeton [Miller et al, 1990] dans le but de répertorier, classifier et mettre en relation de diverses manières le contenu sémantique et lexical de la langue anglaise.

A l'origine, les concepteurs de WordNet l'ont conçue comme une ressource lexicale qui rend compte de l'usage des mots et de leur mise en relation dans la langue [Baziz, 2005]. Ensuite, elle a été perçue comme une représentation conceptuelle (Lexical Conceptuel Graph ou LGC) [Guarino et al, 1999] sous forme ontologie. Dans sa structure, chaque concept « nommé Synset » est composé par un ensemble des termes qui sont reliés par une relation de synonymie. Les concepts sont ensuite reliés par différentes relations sémantiques.

Actuellement, suivant le site de WordNet[8], la version la plus récente (3.0) atteint une taille importante de 12 mégabytes, elle couvre la majorité des noms, verbes, adjectifs et adverbes de la langue anglaise.

2.3.1 Synset WordNet

Le Synset est un ensemble des termes synonymes, il désigne la composante atomique (ou concept) sur laquelle repose WordNet. Autrement dit, un synset dénote un sens qui correspond à un groupe de mot. Par exemple, le nom commun WordNet anglais « car » est défini à l'aide de cinq synsets :

1. car, auto, automobile, machine, motorcar -- *(4-wheeled motor vehicle; usually propelled by an internal combustion engine; he needs a car to get to work)*
2. car, railcar, railway car, railroad car -- *(a wheeled vehicle adapted to the rails of railroad; three cars had jumped the rails)*
3. car, gondola -- *(car suspended from an airship and carrying personnel and cargo and power plant)*
4. car, elevator car -- *(where passengers ride up and down; the car was on the top floor)*
5. cable car, car -- *(a conveyance for passengers or freight on a cable railway; they took a cable car to the top of the mountain)*

Chaque synset dénote une acception différente du mot « car », décrite par une courte définition appelée glossaire. Le premier sens représente le sens le plus courant. Un problème de polysémie est relié à la notion des synsets, qui, dans le contexte d'une phrase ou d'un énoncé sera intéressant de trouver le sens WordNet correspondant à un terme qui sera le plus significatif. Dans ce but plusieurs méthodes de désambiguïsation ont été employées [Agirre et al, 2000] [Agirre et al, 2001] [Banerjee et al, 2003] [Basili et al, 1997] [Dorr et al, 1996] [Fellbaum et al,01] [Karov et al, 1996] [Kwong 2001][Baziz, 2005].

Dans la suite de ce mémoire, nous notons *concept* pour désigner un synset WordNet. Ainsi que, nous utilisons la notion *concept-terme* pour désigner un sens d'un terme appartenant à un concept WordNet.

2.3.2 Relations WordNet

Dans WordNet, les noms et les verbes sont organisés d'une manière hiérarchique. Les relations d'hyperonymie (est-un) et d'hyponymie sont à la base de cette hiérarchie, elles relient les ancêtres des noms et des verbes avec leurs spécialisations. Ces relations permettent de représenter des concepts plus abstraits de plus haut niveau que les mots et leurs sens. Prenons un exemple sur la relation d'hyperonymie du premier sens de mot « car » présenté dans WordNet. Ce concept appartient au synset suivant : « car, auto, automobile, machine, motorcar ». Ainsi, ce synset est relié par une relation hyperonymie à un autre synset WordNet de plus haut niveau « motor vehicule, automotive vehicle » pour obtenir un arbre de concepts de plus en plus généraux reliés par la relation d'hyperonymie:

- car, auto, automobile, machine, motorcar
  - motor vehicle, automotive vehicle
    - vehicle
      - conveyance, transport
        - instrumentality, instrumentation
          - artifact, artefact

8 http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html

- object, physical object
  - entity, something

L'organisation générale de WordNet est sous forme d'ontologie dans laquelle, les concepts sont les synsets et les différentes relations sémantiques permettent de regrouper de manière cohérente les composantes d'un univers linguistique.

Différentes relations sémantiques existent entre concepts. La relation de synonymie permet de relier les termes d'un même synsets. Ensuite, six relations sémantiques de base peuvent exister entre deux synsets (ou expressions) w et w', et être obtenues à partir de WordNet. Ce sont :

- w et w' sont synonymes (w S w') ;

- w figure dans la définition de w', relation dite de glossaire (w G w') ;

- w spécialise w', c.-à-d. que un w "est-un" w' (w I w'); w' I^{-1} w se lit alors w' généralise w;

- w est une partie de w' (w P w'); à l'inverse w' P^{-1} w se lit : w' comporte w comme partie ;

- w et w' sont dans le même domaine thématique (w D w') ;

- w est en relation sémantique avec w' (w R w').

Les relations S, D, et R sont symétriques, tandis que, G, I et P sont antisymétriques. Par ailleurs, S, I et D sont transitives. De plus, il est possible de définir de nouvelles relations à partir de ces relations en prenant leurs unions: w (Ri $\cup$Rj) w' signifie que w est en relation Ri ou Rj avec w'; on peut aussi penser à composer les relations: w RioRj w' ssi $\exists$w°, w Ri w° et w° Rj w', mais cela conduirait à mettre en relation des mots qui sont sémantiquement distants.

2.3.3 Limites de WordNet

WordNet a été utilisé dans un grand nombre d'applications du TAL pour désambigüisation de sens, annotation sémantique de texte, extraction d'information et recherche d'information [Fellbaum, 1998]. Sa structure est en cours de développement, qui, en 2005 son réseau était composé de 1924460 mots ainsi de 109377 concepts [Baziz, 2005] pour atteindre dans sa version actuelle plus de 206941 mots, ainsi que, plus de 117659 synsets (concepts)[9]. Ce qui implique qu'à chaque nouvelle version, le lexique s'enrichit par de nouveaux mots, et des relations sémantiques sont ajoutées, modifiées, ou encore supprimées. En général, suivant les expérimentations, plusieures limites peuvent être soulignées dans l'utilisation de WordNet [De Loupy et al, 2002]. D'une part, elles sont peu adaptées pour traiter les domaines spécialisés qui manipulent un vocabulaire particulier, exemple, un mot non polysémique dans un certain domaine peut être polysémique dans ces bases. D'autre part, la structure même de ces bases peut ne pas répondre d'une manière adéquate à des besoins particuliers, propre à une application visée [Claveau, 2003] ; les relations entre termes de même domaine ne sont pas indiquées systématiquement. Ces limites spécifiques ne peuvent être réglées que par le développement de ressources propres aux domaines et aux besoins.

D'autres limites sont spécifiquement reliées à l'ontologie générée par la relation d'hyperonymie de WordNet. Au niveau catégories lexicales, l'ontologie de WordNet couvre la majorité des noms, verbes, adjectifs et adverbes. Le système de catégorisation pour les noms est complet et précis, les noms sont classés sur plusieurs niveaux d'imbrication où la profondeur pour certaines sections dépasse 10 niveaux. Tandis que, les verbes sont organisés dans un système de classification beaucoup moins élaboré où dans la hiérarchie on passe rapidement d'un concept spécialisé à un concept très général. De nos jours, il n'y a aucune

9 http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html

catégorisation hiérarchique définie pour des adjectifs et des adverbes. Au niveau de la construction des catégories, un déséquilibre est relié précisément à la hiérarchie de la branche nominale où certains mots sont ainsi liés à une grande chaîne de concepts finement gradués, tandis que, d'autres sont très proches des concepts les plus généraux.

2.3.4 WordNet pour d'autres langues que le français

***EuroWordNet*[10]**

EuroWordNet est un réseau sémantique pour plusieurs langues européennes. Chaque langue développe son propre WordNet avec la conception de la base de données et la définition des différentes types de relations, mais ils sont tous interconnectés avec des *liens inter-langues* (*interlingual links*) qui sont enregistrés dans un *Index Inter-langues* (*Interlingual Index ILI*). A partir des Index Inter-Langues, il est possible de trouver des mots dans une langue qui correspondent aux mêmes mots dans n'importe quelle autre langue.

EuroWordNet a produit des WordNet pour les néerlandais, l'italien, l'espagnol, l'allemand, le français, le tchèque et l'estonien. D'autres extensions de EuroWordNet ont ensuite été développées pour plusieurs langues (suédois, norvégien, danois, grec, portugais, basque, catalan, roumain, lithuanien, russe, bulgare et slovène). L'accès à certains échantillons d'EuroWordNet sont disponibles gratuitement, tandis que l'accès à la totalité de la base des données n'est pas libre.

***BalkaNet*[11]**

Une prolonge de la base de données d'EuroWordNet, le projet Balkan WordNet (BalkaNet) a été développé pour d'autres langues européennes : chèque, roumain, grec, turc, bulgare et serbe.

2.4 Modèle de représentation de l'information

Les mots ou termes ou concepts extraits des documents ou d'un texte en générale, sont souvent sous forme de listes (ou ensemble) de mots (ou de concepts). Il existe des représentations encore plus sophistiquées permettant la représentation de l'information sous forme de clusters (ou groupes), de réseaux et de graphes. Nous détaillons dans cette section la représentation sous forme de clusters telle que nous l'avons utilisée dans le cadre de nos travaux. Nous donnerons quelques éléments sur les deux autres modes de représentations par graphes conceptuelles et réseaux sémantiques.

2.4.1 Représentation par clusters

Le partitionnement des données (*groupement* ou *clustering* en anglais) consiste à affecter un ensemble des observations à des sous-ensembles nommés clusters. Les observations dans le même cluster partagent des caractéristiques communes qui correspondent le plus souvent à des critères de proximité [Jain et al, 1988]. Cette similarité est définie en introduisant des mesures et classes de distance entre objets [Candillier, 2006]. Les mesures de similarité sont utilisées pour calculer une mesure de similarité entre deux éléments [Anderberg, 1973]. Dans la littérature, il existe différentes normes pour calculer la distance entre deux points de l'espace [Deza et al, 1994]. Nous pouvons citer, la distance Euclidienne (ou distance norme-2), la distance de Manhattan (ou distance norme-1) [Black, 2006], la norme maximale (ou norme infinie), la distance de Mahalanobis, la distance Hamming, etc.

---

10 http://www.illc.uva.nl/EuroWordNet
11 http://www.ceid.upatras.gr/Blakanet

Pour la construction des clusters, on utilise les méthodes d'apprentissage non-supervisées, ce sont des méthodes statistiques d'analyse de données qui sont utilisées dans plusieurs domaines tels que l'apprentissage automatique (ou *machine-learning* en anglais), la fouille de données (ou *data mining* en anglais), la reconnaissance de formes (ou *pattern recognition* en anglais), l'analyse d'image, le bioinformatique, etc. Dans la Figure 5 nous montrons un exemple de clustering d'un ensemble de carrés qui sont de même couleurs.

Le terme clustering est relié à de nombreux termes qui lui sont similaires, incluant classification automatique, taxonomie numérique et analyse topologique. Il existe de nombreuses applications qui utilisent cette technique, elles sont groupées en deux groupes principaux [Berkhin, 2002] : la segmentation et la classification.

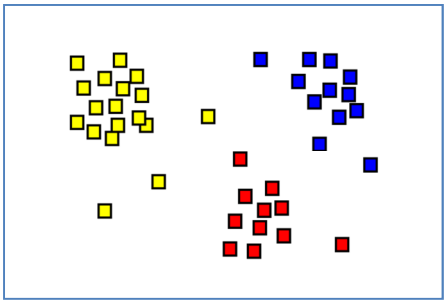


Figure 5: Résultats de clustering pour grouper les carrés par leurs couleurs

La segmentation : consiste à réduire la taille des données à traiter afin de favoriser leur traitement. Cette technique est utilisée par exemple en segmentation d'images, afin d'identifier les différentes zones homogènes de l'esprit décrit (les champs, maisons, routes, fleuves, etc.), ou par exemple cette méthode est utile pour segmenter l'espace dans des bases de données spatiales. Dans d'autres cas, cette méthode est utilisée pour discrétiser une base de données c.à.d. transformer la description complexe des objets par un unique attribut caractérisant leur appartenance à une classe identifiée automatiquement.

La classification : correspond à organiser les connaissances et le travail de chacun au sein de l'ensemble. Le système de classification consiste à classer les objets ou les connaissances, c.à.d. leurs représentation les un(e)s par rapport aux autres. Suivant les objets considérés (les espèces vivantes, les maladies, les produits ou services, les étoiles, les documents dans une bibliothèque...), se sont développés de différents systèmes de classification.

Parmis différentes méhtodes, deux familles de méthodes principales de clustering sont été utilisées dans la RI : les méthodes hiérarchiques et les méthodes par partition [Van Rijsbergen, 1979][Willett, 1988][Rasmussen, 1992]. Les algorithmes hiérarchiques consistent à trouver successivement les clusters en utilisant les clusters déjà construits. Dans ces algorithmes, plus on descend dans la hiérarchie, plus les clusters sont spécifiques à un certain nombre d'objets considérés comme similaires. Par contre, les algorithmes par partition consistent à partitionner l'espace des objets. Ces algorithmes peuvent être utilisés dans les méthodes de clustering hiérarchiques.

Les méthodes de clustering en RI peuvent être utilisés aussi bien pour classer les documents ou encore les termes. Nous présentons dans ce qui suit des travaux proches pour la classification des termes.

2.4.1.1 **Représentation par cohésion lexical calculée par des relations extraites d'un thésaurus**

Différentes tentatives ont été proposées pour améliorer les performances de représentation et d'indexation en exploitant les phénomènes linguistiques [Stokes et al, 2001][Budanitsky, 1999][Kazman et al, 1996][Komineh et al, 1997]. Un des phénomènes linguistiques est la chaîne lexicale, qui crée un lien entre les termes lexicaux reliés dans un document pour représenter la structure de cohésion lexicale du document [Morris et al, 1991]. Si l'on cherche des chaînes lexicales dans l'échantillon du texte de la Figure 6, nous obtenons les six chaînes indiquées dans la droite de la figure. Dans ce schéma, les termes *machine* et *device* sont dans la même chaîne, car ils sont liés par une relation *hyperonyme/hyponyme*. Cette approche montre bien que les termes représentatifs du texte sont les termes *anesthetic* et *machine/device*.

Il est généralement admis que les chaînes lexicales représentent la structure du discours d'un document et elles fournissent les indices sur le sujet d'un document [Morris, 1988][Morris et al, 1991]. Morris et Hirst [Hirst et al, 1998] ont défini une chaîne lexicale comme une chaîne cohérente de mots dans laquelle le critère d'inclusion d'un mot s'appuit sur l'existance d'une relation cohérente entre ce mot et un des mots de la chaîne. Morris et Hirst ont suggéré l'utilisation d'un thésaurus, comme *Roget*, pour spécifier cette notion de relation cohérente entre mots. Les différents types de mot ne sont pas considérés lors de la construction des chaînes lexicales, par exemple pour les pronoms, les prépositions, les auxiliaires verbaux, et les mots très fréquents. Deux mots peuvent être considérés comme liés, s'ils sont liés dans le thésaurus dans un des cinq modes suivants : 1) leurs entrées d'index pointent à la même catégorie du thésaurus ou à des catégories adjacentes. 2) l'entrée d'index d'un mot contient l'autre. 3) l'entrée d'index d'un mot pointe dans une catégorie de thésaurus qui contient l'autre. 4) l'entrée d'index d'un mot pointe vers une catégorie de thésaurus, qui à son tour, cette catégorie est pointée par l'entrée d'index de l'autre. 5) les entrées d'index de chaque mot pointent vers des catégories différentes du thésaurus, qui, à leur tour pointent vers la même catégorie.

Morris et Hirst ont construit et ont évalué les chaînes lexicales manuellement avec cinq textes. Toutefois, ils n'ont jamais été capables d'implémenter une version automatisée de leur algorithme, car les thésaurus en ligne n'étaient pas à leur disposition [Hirst et al, 1998].

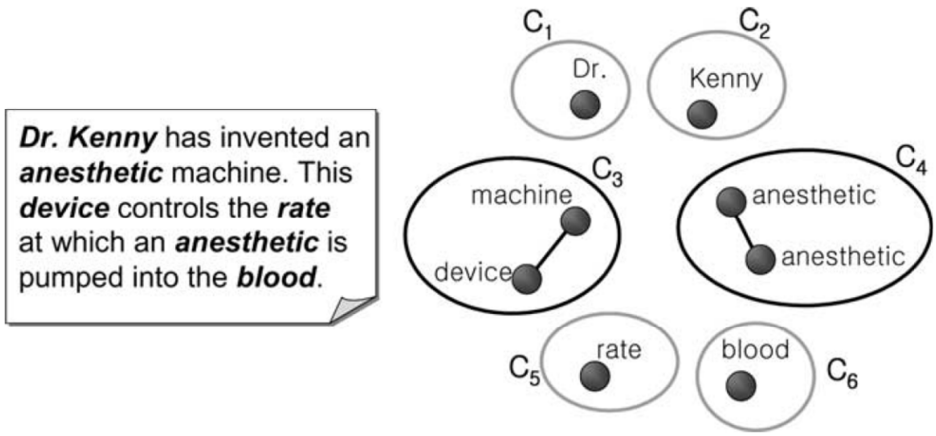


Figure 6: Chaîne lexical d'un simple texte

Dans une tentative de transférer l'algorithme de Morris et Hirst sur les chaînes lexicales à la base de connaissances lexicales en ligne WordNet, Hirst et St_Onge ont défini trois grands types de relations entre les noms de WordNet [Hirst et al, 1998]. Malgré que, WordNet est composé de quatre fichiers : verbes, adverbes, adjectives et noms, mais puisque le fichier de verbes n'a pas de relation avec les trois autres fichiers, et le fichier d'adverbe a une relation indirecte avec le fichier des adjectifs, Hirst et St-Onge limitent le processus de chaînage uniquement à des noms dans leurs recherches. De plus, parce que la structure de WordNet est un peu différente du thésaurus de Roget, ils avaient besoin de remplacer la définition des relations sémantiques définie par Morris et Hirst sur le thésaurus de Roget par des définitions basées sur WordNet, tout en conservant les propriétés essentielles de l'algorithme.

Al-Halimi et Kazman [Kazman et al, 1996][Komineh et al, 1997] ont développé une méthode d'indexation pour les transcriptions des réunions de la conférence par topique en utilisant des arbres lexicaux, la version utilisée est celle de deux dimensions de chaînes lexicales. Ils ont mené une étude préliminaire afin de vérifier l'utilité des arbres lexicaux pour l'indexation automatique de texte arbitraire. Cependant, bien que leur méthode a démontré l'utilité potentielle des arbres lexicaux dans l'indexation et l'extraction de texte, elle est dans sa forme actuelle inappropriée pour une utilisation dans la recherche d'information. Pour qu'une méthode soit utilisable en recherche d'information, chaque terme d'index du document doit avoir un sujet et un poids qui représente le degré d'importance sémantique du terme dans le document. Pourtant, bien que la méthode de Al-Halimi et Kazman peut extraire les sujets comme des termes d'indexes de la transcription d'un séminaire de la conférence, mais elle ne contient pas une fonction pour estimer le poids de chaque sujet extrait.

2.4.2 Représentation sous forme de réseaux sémantiques

Les réseaux sémantiques représentent les premiers modèles à héritage utilisés en représentation des connaissances [Roussey, 2001]. L'origine de ces modèles de représentation est attribué à M Quillian [Quillian, 1968] et donne lieu à de nombreux modèles de représentation des connaissances, citons par exemple, les réseaux sémantiques partitionnés [Hendrix, 1978], les graphes conceptuels [Sowa, 1984], le langage de KL-ONE [Brachman, 1983].

Un réseau sémantique est une structure de graphe orienté et étiqueté, plus précisément un multi-graphe car deux nœuds du graphe peuvent être reliés par plusieurs arcs. Dans un réseau sémantique, il existe deux types de nœuds, les nœuds étiquetés par des constantes de concepts (représentant des catégories taxonomiques représentées par les rectangles jaunes et roses sur la Figure 7), et les nœuds étiquetés par des constantes d'objets (représentant des instanciations des concepts ou des propriétés des concepts représentés par les rectangles verts biseautés et bleus oblongs sur la Figure 7). Et trois types d'arcs connectent les nœuds, les arcs d'agrégation (appelés aussi *"liens isa (IS A)"* - représentés par les liens orange sur la Figure 7), les arcs de composition (appelés aussi "liens hasa (HAS A)" - représentés par les liens mauves sur la Figure 7) et les arcs d'instanciation (appelés aussi : "liens iko (IS A KIND OF)" représentés par les liens verts sur la Figure 7).

Le mécanisme de recherche d'un graphe cible dans le graphe sémantique, correspond à identifier parmi les sous-graphes du graphe, la structure qui correspond au graphe cible. L'utilisation des réseaux sémantiques dans la recherche d'information, correspond à représenter la requête sous forme d'un réseau sémantique dans lequel les connaissances inconnues sont exprimées par des variables. Le raisonnement du système correspond à mettre en correspondance le réseau requête avec une partie du réseau contenant l'ensemble des connaissances afin de déduire les valeurs des variables inconnues.

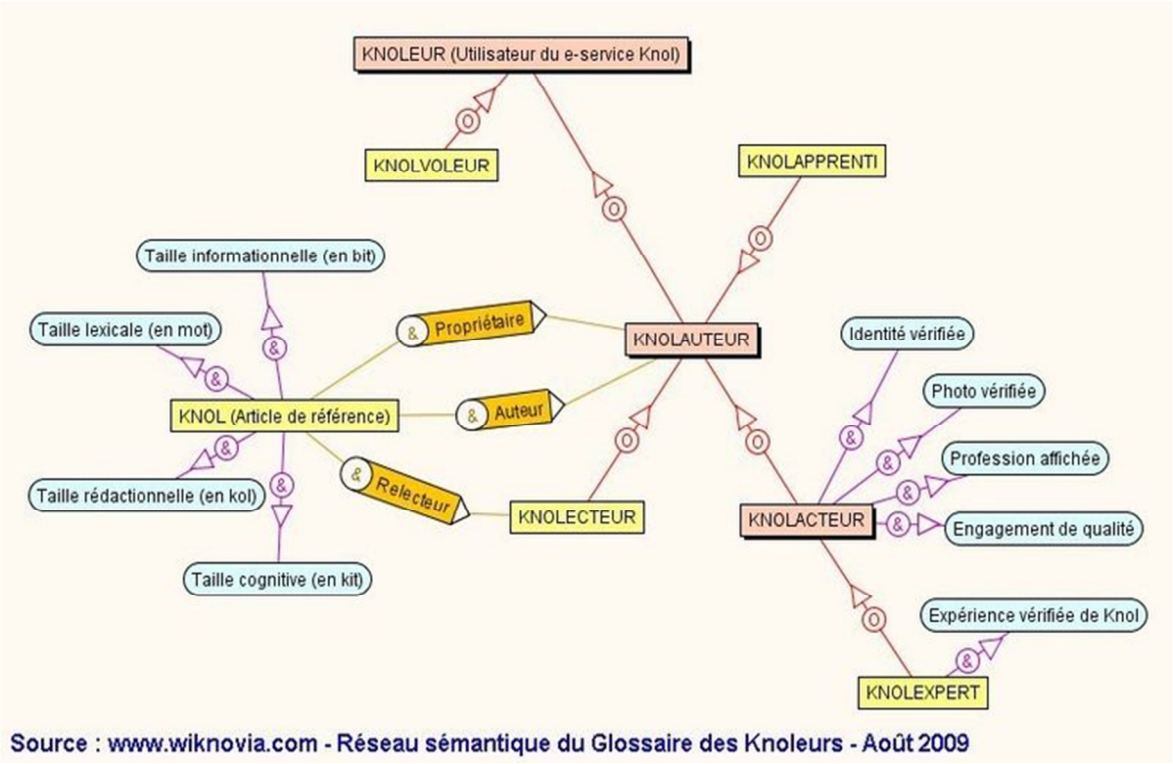


Figure 7: Exemple de réseau sémantique

2.4.3 Représentation sous forme de graphes conceptuels

Les graphes conceptuels (GC) introduits par John F. Sowa depuis 1976 [Sowa, 1984] sont de plus en plus utilisés à l'heure actuelle en Intelligence Artificielle (IA) pour la représentation et le raisonnement sur les connaissances.

Le graphe conceptuel est un multi-graphe [Roussey, 2001], composé de deux sortes de nœuds : les nœuds concepts (ou plutôt concepts), et les nœuds relations (ou les relations) où chacun de ces nœuds a une étiquette. Un nœud concept est étiqueté par un type correspondant à une classe sémantique, et un marqueur précisant une instance particulière de la classe. Un exemple du graphe conceptuel est représenté dans la Figure 8. Cet exemple nous présente un concept étiqueté par 'Automobile : 6793 SY 69'. Ce concept est un nœud *individuel*, qui est une spécification d'un nœud plus générique (ou *concept générique*) qui est étiqueté par 'Automobile'. Il représente une voiture particulière dont le numéro d'immatriculation est '6793 SY 69'. Les relations spécifient les rapports entre les concepts. Dans les graphes conceptuels, un concept est appelé un argument de la relation. Les nœuds relation sont aussi étiquetés par un type.

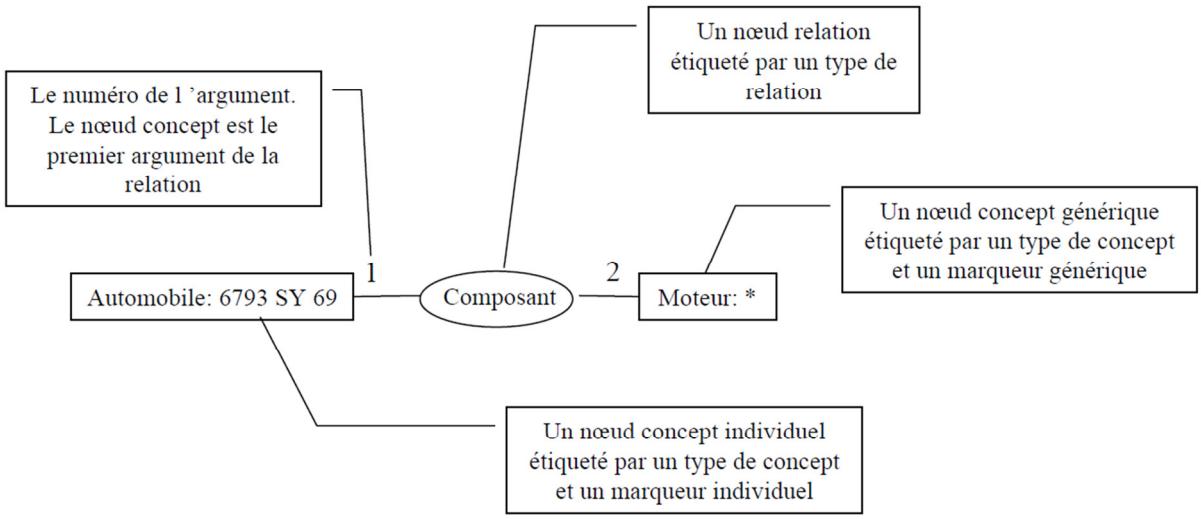


Figure 8: Le graphe conceptuel "une automobile est composée d'un moteur [Roussey, 2001]

Pour la recherche d'information Sowa a proposé un opérateur de projection '≤' qui permet de comparer entre les graphes (exemple pour deux graphes G et H, on peut dire que '$G \leq H$' si G est une spécialisation de H). Cet opérateur permet de détecter si un graphe '*G*' est une spécification ou une généralisation d'un autre graphe '*H*'. Pour que cet opérateur soit appliqué sur la recherche d'information [Chevallet, 1992][Van Rijsbergen, 1986], Sowa a défini un opérateur qui permet de transformer un graphe conceptuel en une formule logique du premier ordre.

Les approches concepuelles ont été largement utilisées dans la recherche d'information. Elles tentent d'identifier et d'extraire les concepts apparaissant dans les documents. Ces concepts sont souvent extraits à partir d'une ressource externe (dictionnaire, thésaurus, ontologie). Dans [Montes-y-Gómez et al, 2000] et [Thomopoulos et al, 2003] l'utilisation de graphes conceptuels pour représenter des documents (et des requêtes) est discutée. Les auteurs proposent une méthode pour mesurer la similarité de phrases représentées par des graphes conceptuels. Gonzalo et al. [Gonzalo et al, 2004] proposent une méthode de représentation des documents basée sur les « synsets » de WordNet. Le modèle vectoriel est employé, en utilisant les synsets comme espace d'indexation au lieu de mots. Dans un esprit similaire, Richardson et Smeaton [Richardson et al, 1994] ont proposé de représenter les documents et les requêtes sur la base des noms de concepts extraits de WordNet. Baziz et al. [Baziz et al, 2004] ont proposé un modèle plus général de RI à base de concepts, qui représente les documents et les requêtes comme des sous-arbres de concepts (noeuds) extraits d'une ontologie. Les représentations du document et de la requête sont non seulement des ensembles de concepts apparaissant dans leurs contenus, mais ils sont de plus complétés par des concepts (noeuds) intermédiaires. Roussey et al. [Roussey et al, 2001] ont proposé une extension du modèles des Graphes conceptuels de Sowa permettant d'améliorer la description sémantique des documents dans un contexte multilingue tenant compte de l'existance de plusieurs vocabulaire, un par langue. Ces vocabulaires constituent les connaissances terminologiques du modèle, à différencier des connaissances propres au domaine d'étude.

2.4.4 Autres modèles de représentations des connaissances

D'autres modèles de représentation des connaissances existent tels que les langages de frames qui regroupent les connaissances en paquets. Dans ce modèle, un frame représente l'unité sémantique (ou le concept) et regroupe en même temps le concept et sa description. Le

premier modèle de frame est proposé par [Minsky, 1975], dans ce modèle le frame correspond à une structure dynamique représentant des situations prototypiques [Roussey, 2001]. Autres modèles telles que les logiques de description représentent un formalisme de représentation des connaissances très répandu actuellement. KL-ONE [Brachman et al, 1985] développé en 1978 est considéré comme le précurseur des logiques de description.

2.5 Résumé d'un texte ou d'un document

Le résumé d'un texte ou d'un document permet de le présenter de différentes manières, citons par exemple : sac de mots clés (*BOW, Bag Of Word*), Nuage de mots clés (*Tag Cloud*) ou par liste des phrases extraites automatiquement. La représentation par sac de mots clés permet de représenter un document ou un texte par une liste de mots non ordonnés qui le compose [Salton et al, 1988]. Dans ce mode d'affichage, on ne peut pas distinguer les mots qui sont plus significatifs pour comprendre le contenu d'un texte, d'où la nécéssité d'autre mode qui représente l'importance des termes dans un texte. La plupart des modes consistent à représenter un texte comme un sac de mots pondérés [Kraft et al, 1999] [Hisamitsu et al, 2003]. La pondération estime le degré d'importance d'un mot pour décrire le contenu d'un texte.

La représentation par nuage de mots clés permet de visualiser un document ou un texte par un nuage de mots-clés (ou tag cloud) qui représente les mots les plus fréquents d'une manière similaire à celle présnetée dans la Figure 9. Concrètement, plus un mot est cité dans un texte, plus il apparaît en gros dans le nuage de mots-clés. En général, les mots clés sont des mots simples et typiquement listés par ordre alphabétique [Halvey et al, 2007]. Chacun des mots clés dans un nuage réfère une collection des articles qui lui sont associés. Historiquement, le site de partage de photos *« Flickr »* fut le premier à implémenter ce système de représentation, créé par Stewart Butterfield [Bausch et al, 2006]. Les nuages de mots clés sont popularisés par les sites de *« Del.icio.us »* et *« Technorati »*. Dans la Figure 9, nous présentons un exemple de nuages des mots clés comme représenté par le site de *Flickr*[12] où chaque mot-clé permet de trouver tous les photos et les vidéos qui ont un point commun. Il existe deux grandes familles de nuages de mots-clés. C'est plus par leur valeur sémantique que par leur apparence que l'on distingue ces catégories. La première famille de nuage de mots-clés classe les concepts selon le critère de la répétition d'un mot dans un article [Bielenberg et al, 2006]. Il s'agit donc d'une méta-donnée permettant de symboliser par ordre d'importance les concepts que recouvre l'article en cours. La seconde, plus transversale, regroupe en nuage de mots-clés, les mots-clés revenant le plus souvent dans un site ou dans un annuaire de sites. Il s'agit donc de mettre en avant la popularité d'un concept, qui a fédéré plusieurs rattachements dans un site ou un ensemble de sites. Cela est particulièrement util à une navigation transversale, permettant de balayer l'intégralité du contenu d'un site à travers le fil conducteur du mot-clé auquel on s'intéresse. Dans le cas d'un annuaire de site tel que Flickr ou Technorati, le nuage de mots-clés permet alors de mesurer d'un coup d'œil les tendances du moment à travers les termes revenant le plus souvent dans les sites syndiqués. Les mots dans un nuage de mots clés peuvent apparaître dans un ordre alphabétique, ou dans un ordre aléatoire, ainsi que listés par leurs poids, etc. Certains préfèrent de grouper sémantiquement les mots clés par des clusters [Hassan-Montero et al, 2006] [Kaser et al, 2007] d'une manière que les mots clés similaires apparaissent proches les unes des autres.

12 http://www.flickr.com/photos/tags/

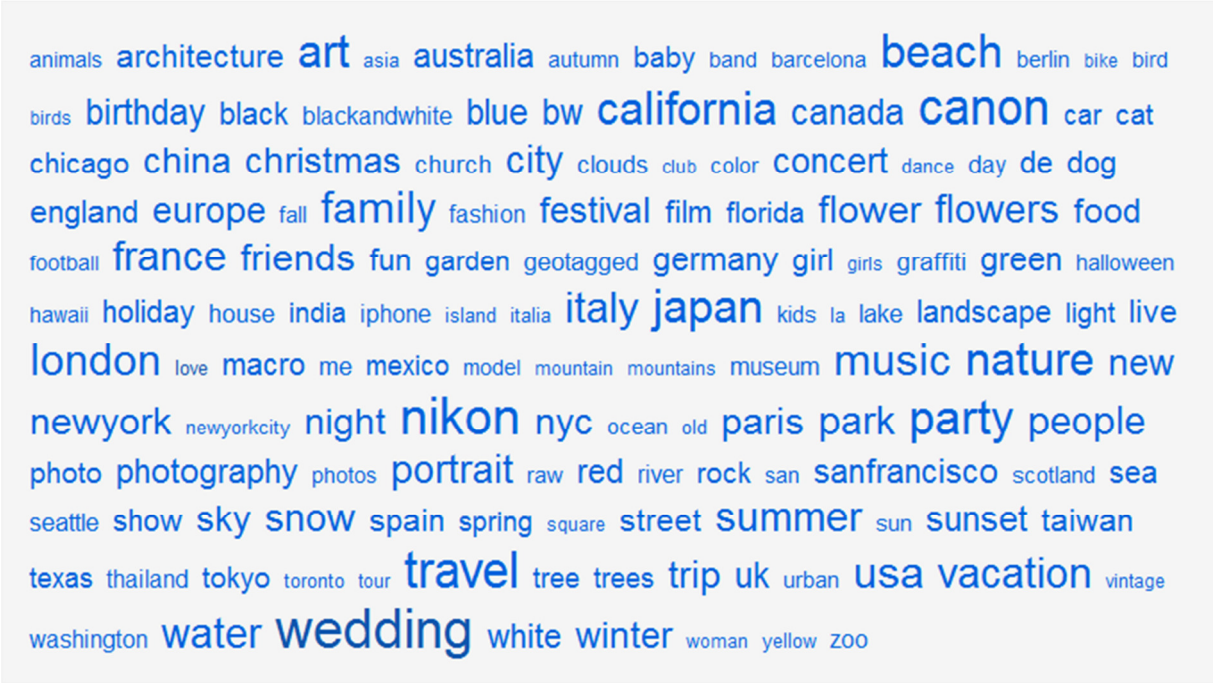


Figure 9: Exemple sur les nuages des mots clés

La représentation par des phrases extraites automatiquement fait partie de résumés automatiques par extraction. Comme nous l'avons évoqué dans l'introduction de ce chapitre, les méhtodes de résumés automatiques sont groupées en deux familles : extraction et abstraction. Nous nous n'intéressons qu'aux approches de résumé par extraction qui sont en relation avec nos travaux. Comparées aux approches de resumés par extraction, les approches par abstraction consistent à résumer un texte par un ensemble des unités (paragraphes, phrase, mots, etc.) qui ne sont pas forcément présentes dans le document source. Quant à eux, les approches par extraction consistent à selectionner les informations essentielles présentes dans un document (mots, phrases, etc.), ainsi qu'à produire un extrait par assemblage de ces derniers. Un extrait, comme son nom l'indique, est une partie extraite d'un document source visant à donner un aperçu de son contenu. Lin et Hovy [Lin et Hovy, 2003] ont remarqué qu'environ 70% des phrases utilisées dans les résumés créés manuellement sont retirés de texte source sans aucune modification. Les approches d'extraction des phrases sont les plus utilisées de nos jours. Exemple de l'utilisation des phrases extraites automatiquement, les « synopsis Google » ou « snippet Google en anglais » employés par le moteur de recherche Google pour représenter la liste des résumés des pages web en réponse à une requête utilisateur, voir Figure 10: Snippet Google. Dans ces snippets, quelques phrases compressées sont extraites d'une page web et qui contiennent les termes de la requête utilisateure, ainsi qu'un lien vers la page contenant ces résumés.

Figure 10: Snippet Google

Dans cette section, nous représentons différentes approches d'extraction automatiques des phrases pour la production des résumés automatiques: les approches classiques, les approches par apprentissage, les approches basées sur le centroïde, les approches exploitant la structure rhétorique et celles basées sur les graphes

2.5.1 Les approches classiques

Luhn [Luhn, 1958], est parmi les premiers qui ont travaillé sur les résumés automatiques. Il a décrit une technique spécifique aux articles scientifiques basée sur la distribution des fréquences des mots dans le document pour pondérer les phrases. Dans son article, il a décrit quelques avantages que présentent les résumés produits de manière automatique par rapport aux résumés manuels : coût de production très réduit, non soumis aux problèmes de subjectivité et de variabilité observés sur les résumés effectués par des professionnels. Dans sa technique, il a proposé une version primitive de lemmatisation « *stemming* » pour s'affranchir les variations de l'orthographiques des mots en regroupant les mots porteurs du même sens.

D'autres systèmes influencés par l'idée de Luhn, sont également basés sur des techniques statistiques pour la production automatique de résumés. Edmundson [Edmundson, 1969] a étandu les travaux de Luhn en tenant compte de la position des phrases, des mots présents dans les différentes zones d'un document (titres, sous-titres, etc.) et de la présence de mots indices (exemple *significant*, *impossible*, *hardly*, etc.). Pour évaluer son approche, il a comparé manuellement les résumés produits par son système à des résumés de référence (phrases extraites manuellement). Edmundson a pu montrer que la combinaison – position, mots de titres, mots indices- était plus performante que la distribution des fréquences de mots. Il a également trouvé que la position de la phrase dans le document était le paramètre le plus important.

Une validation des approches d'extraction automatiques des phrases a été menée par les recherches de Pollock et Zamora [Pollock et Zamora, 1975] par la production des résumés à partir d'articles scientifiques de Chimie au sein du *Chemical Abstract Service* (CAS). Dans sa proposition, il introduit pour la première fois un processus de nettoyage des résumés des phrases en se reposant sur des opérations d'élimination. Les phrases commençant par exemple par « in » (exemple « in conclusion ») ou finissant par « that » sont éliminées du résumé. Afin que les résumés satisfassent les standards imposés par le CAS, une normalisation du vocabulaire est effectuée, elle inclut le remplacement des mots/phrases par leurs abréviations, une standardisation des variantes orthographiques (exemple conversion de l'anglais UK en anglais US) et le remplacement des noms de substances chimiques par leurs formules.

Ces travaux sont à la base du résumé automatique de textes. De leur analyse émerge une méthodologie de production des résumés en deux étapes. La première consiste à une analyse du document, cette étape correspond à la compréhension, la sélection et l'extraction d'unités (généralement phrases) importantes. La seconde étape consiste à une génération d'un résumé par assemblage des unités importantes, dans cette étape des techniques de reformulation, de généralisation ou de compression sont employés.

2.5.2 Les approches par apprentissage

L'importance de la phrase, comme nous l'avons présentée dans la section précédente, est déterminée en fonction de certains paramètres tels que la position de la phrase ou la présence de certains mots. Selon le type de document traité, ces paramètres sont utilisés dans les résumés pour extraire les phrases significatives. Exemple, dans le cas d'articles journalistes, les premières phrases sont souvent les plus importantes [Brandow et al, 1995] tandis que pour les articles scientifiques, les phrases provenant de la conclusion seront privilégiées. C'est dans ces cas que les approches par apprentissage seront intéressantes. Dans l'apprentissage, chaque paramètre est estimé en comptant son occurrence dans un corpus. Ainsi, de nombreuses recherches ont tenté d'analyser comment un corpus composé de paires [document/résumé associé généré manuellement] peut être utilisé afin d'apprendre automatiquement des règles ou des techniques pour la génération de résumé.

Une méthode d'apprentissage à partir d'un ensemble de données dérivée de [Edmundson, 1969] est décrite dans les travaux de [Kupiec et al, 1995]. Dans cette méthode, un classifieur bayésien entraîné sur un corpus de 188 paires [document/résumé], calcule pour chaque phrase une probabilité qu'elle soit incluse dans le résumé. Notons s une phrase, S l'ensemble des phrases qui composent le résumé et $F_1,\ldots, F_k$ les paramètres. En supposant que les paramètres sont indépendants :

$$P(s \in S|F_1, \ldots, F_k) = \frac{\prod_{i=1}^{k} P(F_i|s \in S).P(s \in S)}{\prod_{i=1}^{k} P(F_i)}$$

Les paramètres $F_1,\ldots, F_k$ sont ceux décrits par [Edmundson, 1969] en plus des deux paramètres supplémentaires : la longueur de la phrase et la présence des mots en majuscules. Aone et al. [Aone et al, 1999] ont étendu cette approche en utilisant des paramètres plus riches comme la présence de *signature word* (mots indiquant les concepts clés d'un document). Dans les apprentissages, on note que la taille de l'ensemble de données utilisées est très faible.

Lin [Lin, 1999] de sa part, a modélisé la problématique d'extraction de phrases avec des arbres de décision. Dans sa proposition, il a mis fin à l'hypothèse d'indépendance des paramètres. Plusieurs combinaisons des paramètres ont été évaluées, exemple l'utilisation du paramètre de position seul ou l'addition des valeurs de tous les paramètres. L'évaluation consiste en l'appariement des phrases extraites par le système avec celles extraites manuellement. La simple combinaison des paramètres a été meilleure que le classifieur par arbre de décision sur trois thématiques malgré que les arbres de décision s'apparaissent être globalement plus performants. Suite à ces travaux, Lin a conclu que certains des paramètres étaient indépendants vis-à-vis des autres.

D'autres approches par apprentissage ont été expérimentées, exemple celle de Conroy et O'leary [Conroy et O'leary, 2001] utilisant le Modèle de Markov Cachés (HMM, *Hiden Markov Model*) ou celle de Svore et al. [Svore et al, 2007] utilisant les réseaux de neurones. Malgré la performance de ces approches d'extraction par apprentissage, il n'y a pas de garantie que les résumés puissent être exploités. En effet, les phrases sélectionnées peuvent ne

pas être cohésives, exemples les anaphores (Par exemple dans le texte : « La baleine bleue est en voie de disparition, elle est pourtant le plus grand mammifère marin. », le pronom « elle » est une anaphore se rapportant à « la baleine bleue ».) non résolue. De plus, les paramètres utilisés lors de l'apprentissage peuvent ne pas être compatibles pour les différents types de documents.

2.5.3 Les approches basées sur le centroïde

Ces approches consistent à choisir les phrases les plus centrales dans un cluster (ensemble de documents), celles qui donnent les informations nécessaires et suffisantes en relation au thème principal du cluster. La centralité d'une phrase est généralement définie en fonction de la centralité des termes qu'elle contient. D'une manière générale, pour évaluer la centralité d'un mot nous regardons le centroïde correspondant au cluster de documents dans un espace vectoriel. Le centroïde d'un cluster est un pseudo-document constitué des mots ayant un score tf×idf supérieur à un seuil prédéfini. Dans les résumés basés sur le centroïde [Radev et al, 2000], les phrases contenant plusieurs mots de centroïde sont considérées centrales. C'est une mesure qui permet d'estimer si une phrase est proche du centroïde du cluster. Les approches de resumé basées sur le centroïde ont déjà donné des résultats prometteuses, ainsi que des résultats dans les premiers systèmes de resumé multi-document basés sur le Web [Radev et al, 2001]

2.5.4 Les approches exploitant la structure rhétorique

D'après la théorie de la structure rhétorique (RST, *Rhetorical Structure Theory*) [Mann et Thompson, 1988], un texte peut être organisé en éléments reliés entre eux par des relations. Les relations dans cette représentation peuvent être de deux types : des *satellites* ou des *noyaux*. L'idée qu'un satellite a besoin d'un noyau pour être compréhensible tandis que la réciproque n'est pas vraie. De plus, des schémas sont utilisés pour spécifier la composition structurale du texte. Le schéma le plus fréquent est constitué de deux segments de textes (phrases, clauses, etc.) reliés de telle sorte que l'un des deux joue un rôle spécifique par rapport à l'autre, par exemple une déclaration suivie d'une démonstration l'appuyant. La RST postule une relation de démonstration entre les deux segments et considère la déclaration comme plus importante (noyau) pour le texte que la démonstration (satellite de ce noyau). Cette notion d'importance, nommée « nucléarité », est centrale pour la RST. Ces représentations ainsi construites peuvent être utilisées pour déterminer les segments les plus importants du texte. Ces idées sont utilisées dans des systèmes visant à produire des résumés [Ono et al, 1994][Marcu, 1997]. Dans ces approches, ils construisent un arbre rhétorique basé sur la présence de marqueurs explicites dans le texte. Dans [Ono et al, 1994], la phrase est l'unité minimale de l'analyse. Selon le rôle rhétorique des phrases, elles sont pénalisées dans l'arbre. Un poids « 0 » est attribué à chaque phrase noyau et un poids de « 1 » à chaque phrase satellite. Le poids d'une phrase est calculé par la somme des poids qui la sépare de la racine. Dans [Marcu, 1997], la clause est l'unité minimale. Ces unités sont poussées de manière récursive, une clause au niveau *n* de l'arbre est poussée au niveau *n-1* si elle est le noyau de la relation au niveau *n-1*.

La production des résumés basée sur la RST donne des résultats prometteurs [da Cunha et al, 2007]. Un réel problème est relié à l'abscence d'étiquetteur automatique adéquat (mais également indépendant de la langue) qui permettrait d'identifier la composition structurale des documents. L'evaluation de ses méthodes repose sur une annotation entièrement manuelle des structures RST dans les documents.

2.5.5 Les approches basées sur les graphes

Les approches par analyse de graphes ont été utilisées avec succès dans les réseaux sociaux, l'analyse du nombre de citations ou de la structure du Web. Les algorithmes de pondération tels que HITS [Kleinberg, 1999] ou PageRank [Brin et Page, 1998] peuvent être vus comme les éléments clés dans le domaine de la recherche sur internet, plus précisément le classement des pages Web par l'analyse de leurs positions dans le réseau et non pas de leurs contenus. Ces algorithmes permettent de décider de l'importance du sommet d'un graphe en se basant sur une analyse récursive du graphe complet. Les approches proposées par [Erkan et Radev, 2004][Mihalcea, 2004] reposent sur ce concept pour la pondération des phrases. Ainsi, un document est représenté par un graphe de phrases reliées par des relations issues de calcul de similarité. Les phrases extraites sont choisies suivant des critères de centralité ou d'importance dans le graphe, puis assemblées pour produire un résumé.

2.6 Conclusion

Le but dans ce chapitre est de présenter les différentes méthodes de représentation d'un document. Nous avons abordé les différentes méthodes d'identification des descripteurs à partir des documents ainsi que les différentes formes que peuvent prendre un descripteur de document. Ensuite, nous avons décrit les différentes approches possibles statistiques, linguistiques, hybrides, sémantiques et conceptuelles pour l'extraction de ces descripteurs. De plus, nous avons présenté les différentes formes de représentation de l'information sous forme de clusters, réseaux sémantiques, réseaux conceptuels, graphes, etc. et nous avons détaillé les différents domaines d'utilisation des méthodes de clustering sur lesquelles nous basons notre proposition de la thèse. Ainsi, nous avons présenté les différents modes d'affichage du contenu d'un document par sac de mots clés, nuage de mots clés ou par des phrases extraites dans le cadre de résumé automatique des textes. Les différentes approches d'extraction des phrases ont été présentées.

PARTIE II : **CONTRIBUTION**

Chapitre 3 Extraction de termes significatifs d'un document

3.1 Introduction

Nous décrivons dans ce chapitre l'approche que nous proposons pour l'extraction des termes significatifs d'un document. La démarche que nous avons définie est réalisée en deux étapes. La première consiste à extraire les concepts correspondant aux termes représentatifs d'un document. La deuxième étape consiste à regrouper les concepts sous forme de clusters. Le cluster se réfère à un même concept (thème). Ce regroupement est effectué en s'appuyant sur les relations sémantiques définies dans les ontologies. De plus, nous définissons plusieurs facteurs, tels que, taille d'un cluster, son poids, centralité d'un concept-terme ou sa spécificité, dans le but de mieux caractériser, d'une part les concepts-termes appartenant à un cluster, et d'autre part les clusters entre eux. Ces paramètres sont au cœur de nos contributions présentées dans les chapitres 4 et 5.

Ce chapitre est organisé de la façon suivante : nous présentons tout d'abord nos motivations. Ensuite dans la section suivante, nous décrivons la méthode d'extraction des concepts en se basant sur l'ontologie de WordNet. Dans la section suivante, nous détaillons notre méthode de représentation d'un document par 'clusters conceptuels' et comment nous l'utilisons pour définir des critères de pondération des termes d'un document (tels que la taille du cluster, son poids qui est la somme des fréquences des mots composant ce document, mais aussi la spécificité de ces mots, ou leur centralité dans les "clusters", etc.). Nous finissons par une conclusion.

3.2 Extraction de termes significatifs d'un texte

En général, si nous souhaitons comprendre le contenu d'un document, l'étape essentielle consiste à extraire ses différents concepts. Cette étape consiste à déterminer pour chaque terme identifié dans le document un concept qui lui correspond dans une ontologie *O* des concepts, qui lui est supposée suffisante pour décrire le contenu du document considéré avec un certain degré de précision. Ces concepts peuvent correspondre à des mots simples ou à des groupes de mots (*multi-terme*) extraits du document.

Bien que plusieurs ontologies puissent être utilisées, nous avons décidé de tester notre approche en utilisant WordNet comme ontologie. WordNet qui, en plus de sa bonne couverture du lexique de la langue anglaise, est composée des concepts (nommés *synset*) organisés d'une manière hiérarchique par une relation *est-un* (ou *is-a* en anglais). En plus, WordNet fournit un ensemble de relations sémantiques permettant de mettre en relation ces différents concepts pour prendre la forme d'une ontologie [Guarino et al, 1999].

Etant convaincu que les chaînes lexicales offrent une bonne représentation des structures des textes [Morris, 1988][Morris et al, 1991], où chaque chaîne représente une idée conceptuelle représentée dans le texte, dans notre proposition nous considérons une chaîne lexicale comme un 'cluster conceptuel' composé d'un ensemble de concepts de WordNet reliés par les différents liens offerts par son ontologie.

L'approche que nous proposons s'appuit sur trois étapes. La première étape consiste à extraire les concepts de l'ontologie qui sont attachés aux documents. La deuxième consiste à

grouper les concepts qui sont en relation en clusters. La dernière étape consiste à caractériser le cluster selon plusieurs facteurs.

Dans la suite, nous utilisons la notion concept-terme comme nous l'avons définie dans la section 2.3.1 pour désigner un sens d'un terme appartenant à un synset WordNet.

3.2.1 Extraction des concepts

Dans cette section, nous souhaitons extraire à partir d'un texte une liste des concepts-termes WordNet désambiguïsés qui représentent son contenu. Les concepts-termes peuvent correspondre à des termes simples ou composés (multi-termes) du texte. Pour un terme *t*, Le *concept-terme* comme nous l'avons présenté précédemment désigne un sens de *t* arppartenant à un synset (concept) WordNet.

Nous reprenons le même principe de detection des termes simples et composés (multi-termes) proposé par Baziz et présenté dans la section 2.2.6.1. Ensuite, nous représentons chaque terme par son concept-terme qui lui correspond dans WordNet. La détection du meilleur concept-terme représentant le meilleur sens d'un terme suivant son contexte d'apparition nécessite l'utilisation d'une méthode de désambiguisation. Dans un premier temps, nous utilisons la méthode proposée par Baziz détaillée dans 2.2.6.1, ensuite, nous proposons une nouvelle méthode de désambiguïsation basée sur la notion de centralité. Cette méthode est décrite dans le chapitre 5.

Dans notre méthode nous ne gardons que les noms (WordNet est composé de quatre fichiers : verbes, adverbes, adjectives et noms) mais le fichier de verbes n'a pas de relation avec les trois autres fichiers. et le fichier d'adverbe a une relation indirecte avec le fichier des adjectifs. En fait, les relations sémantiques existantes entre les verbes, les adjectifs et les adverbes ne sont pas bien élaborées, voir à ce sujet les travaux de [Fellbaum et al, 1998][Budanitsky, 1999].

Résumons cette étape, l'extraction des concepts d'un texte donné consiste dans une première étape à extraire ces termes simple et composé (multi-terme). Pour chacun des termes extraits nous souhaitons le rattacher à un unique concept-terme de WordNet. Dans une deuxième étape, nous identifions pour chaque terme les différents concepts-termes (appartenant à des concepts (synsets) différentes dans l'ontologie) susceptibles de représenter son sens dans le document. Dans une dernière étape, nous employons une méthode de désambiguïsation pour choisir à partir des concepts-termes choisis, un unique concept-terme représentant le meilleur sens du terme.

3.2.2 Grouper les termes reliés

Dans la section précédente nous avons présenté la méthode d'extraction des concepts à partir d'un texte. Dans cette section, nous présentons la méthode de construction des clusters en groupant dans un même cluster les concepts d'un document qui sont en relation. Nous décomposons cette méthode en deux étapes, la première étape consiste à extraire les relations existantes entre les concepts-termes d'un texte. La deuxième consiste à grouper les concepts-termes qui sont en relation dans un même groupe nommé « *cluster* ».

Plus en détail, dans la première étape nous identifions les relations existantes entre deux concepts-termes (on ne s'occupe que des groupes nominaux) d'un texte. Dans notre proposition nous nous basons sur WordNet [Miller et al, 1990]. Nous n'utilisons pas toutes les relations proposées, nous utilisons uniquement les relations de type synonymie, est-un, est-une-partie-de, même-domaine et dérivé qui sont désignées respectivement pas S, I, P, D et R et leurs relations inverses si elles existent.

Nous considérons que deux concepts-termes sont en relation dans deux cas. Dans le premier cas, ils peuvent être directement reliés par l'une des relations S, I, P, D, R noté par X où $X = S \cup I \cup I^{-1} \cup P \cup P^{-1} \cup D \cup R$. Dans le deuxième cas, ils peuvent être indirectement reliés par une chaine de X-relation à travers un concept intermediaire qui n'existe pas dans le texte, ce cas est applicable sur les relations transitives, comme la relation *est-un*. Dans le cas d'une relation *est-un*, deux concepts présents dans un texte peuvent être présents en relation s'ils ont un ancêtre commun (non présent dans le texte) dans la hiérarchie de l'ontologie. Prenons un exemple sur les relations indirectes, les deux concepts-termes « *ambulance* » et « *Jeep* » s'ils apparaîssent dans un même document, ces deux concepts-termes n'ont pas une relation directe dans WordNet, mais ils sont indirectement reliés à travers un ancêtre (concept WordNet) commun « *car* » à travers la relation d'*hypernyms* « est-un ». Dans le cas des concepts isolés qui n'ont pas un ancêtre commun avec d'autres concepts-termes du texte doivent rester isolés.

Dans la deuxième étape, nous regroupons les concepts qui sont en relation dans un même « cluster » noté C_k. Formellement, un cluster qui n'est pas un singleton peut avoir un chemin entre n'importe quelle paire de concepts de cluster, constituée d'une séquence de concepts X-reliés: on recherche ainsi, les parties connexes du graphe créées par X. L'obtention des clusters est la partie la plus coûteuse de l'approche, mais le calcul de l'ensemble X(w) des mots en relation X avec w peut se faire hors ligne pour chaque mot w. Nous définissons C_k par :

$$C_k = \{t_j \,/\, \forall\, t_i \,\&\, t_j \in\ C_k\ , \exists\ SX\ tel\ que\ t_i\ SX\ t_j\}$$

Avec t_i et t_j sont deux concepts-termes désambigüisés de l'ontologie, et *SX* représente un chemin entre t_i et t_j constitué d'une séquence de concepts du C_k X-reliés.

Illustrons cette méthode par un petit exemple. Nous souhaitons représenter la phrase suivante par cluster: « ***Compensation*** *is available to any* ***individual***, ***business***, *private* ***organisation*** *or public* ***body*** … ». Premièrement, nous extractons à partir de cette phrase les termes qui correspondent à une entrée dans WordNet que nous marquons en gras. Ensuite, nous identifions pour chacun de ces termes un unique concept-terme qui lui correspond. Enfin nous groupons les concepts en relation dans des clusters comme présenté dans la Figure 11.

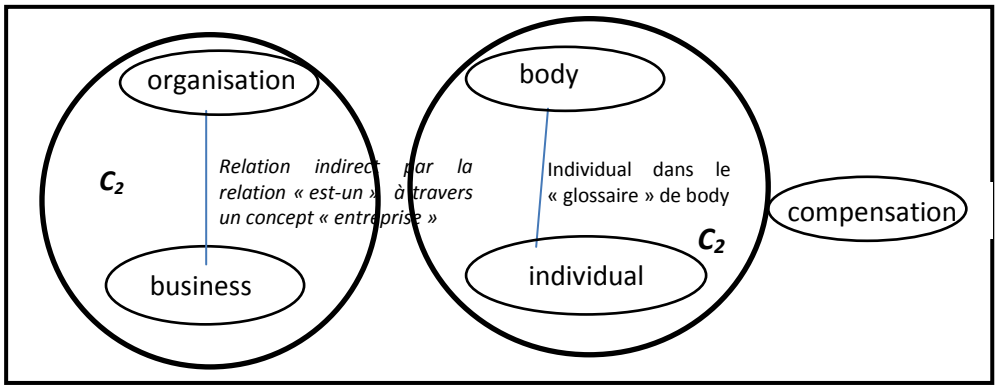


Figure 11: Exemple de création des clusters

Dans la Figure 11 les concepts « *organisation* » et « *business* » sont indirectement reliés à travers un concept « *entreprise* » qui n'existe pas dans la phrase pour former un cluster C_1. Ainsi les concepts « *body* » et « *individuel* » sont reliés par une relation glossaire pour former

un cluster C_2. Le concept « *compensation* » n'a aucun relation avec les concepts de la phrase et reste isolé.

3.2.3 Facteurs pour caractériser les clusters et leurs concepts

Afin de mieux caractériser les clusters et leurs contenus (concepts-termes), nous proposons plusieurs facteurs pour quantifier l'impact d'un cluster vu d'un concept-terme.

3.2.3.1 **Centralité d'un concept-terme**

A l'intérieur de chaque cluster d'un document *d*, chaque concept-terme *t* a, en effet, un niveau de "centralité", fonction du nombre de mots du cluster avec lequel il est en relation directe. Formellement, la centralité $C_{t,d}$ d'un concept *t* dans un document *d* est estimée par le nombre de concepts de *d* qui sont en relation avec *t* :

$$c_{t,d} = \#R(t, t_k) \; \forall \; t_k \; \in d$$

Avec $\#R(t, t_k)$ est le nombre des relations existantes entre un concept *t* et les autres concepts t_k du document.

3.2.3.2 **Fréquence d'un concept-terme**

La fréquence d'un concept-terme est évaluée par son nombre d'occurrences dans le texte (et elle est éventuellement normalisée), les concept-termes se trouvent dans le même "synset" (mots synonymes au sens de la relation S) étant comptabilisés ensemble. Formellement, la fréquence d'un concept-terme *t* dans le document *d* est notée par $f_{t,d}$ qui est égale à la somme des fréquences $f_{ti,d}$ de chaque concept t_i dans le synset de *t* noté *synset(t)* dans le document *d*.

$$f_{t,d} = \sum f_{ti,d} \; \forall \; t_i \; \in synset(t)$$

3.2.3.3 **Fréquence conceptuelle**

La majorité des techniques de calcul de la fréquence des mots se base sur un comptage du nombre de fois que ce mot apparaît effectivement dans le texte. De ce fait, un terme t_1 qui apparaît une seule fois dans un document comportant des termes qui lui sont sémantiquement proches, peut avoir une fréquence plus faible qu'un terme t_2 qui apparaît 2 fois mais qui n'a aucune relation avec les autres termes du document. Afin de « corriger » cette fréquence, nous proposons ce que nous appelons la fréquence conceptuelle. La fréquence conceptuelle d'un concept-terme *t* dans un document *d* est estimée par la somme des fréquences des concept-termes t_k qui sont X-reliés à *t* noté par :

$$Cf_{t,d} = f_{t,d} + \sum\nolimits_{R(t,t_i)} f_{ti,d} \; \forall \; t_i \; \in d$$

Avec $R(t, t_i)$ signifie qu'il existe une relation X entre *t* et t_i.

Exemple, pour le concept-terme *individual* présent dans la Figure 11 extrait de la phrase « ***Compensation*** *is available to any* ***individual***, ***business***, *private* ***organisation*** *or public* ***body*** … » que nous notons *d*. Sa fréquence normale $f_{individual,d}$= 1, comparant à sa fréquence conceptuelle $Cf_{individual,d}$ qui est égale à :

$Cf_{individual,d} = f_{individual,d} + f_{body,d}$ = 1 + 1 = 2.

Nous pouvons constater que la *fréquence conceptuelle* de *individual* est plus significative que sa propre fréquence classique.

3.2.3.4 **Spécificité**

La spécificité S_t d'un terme *t* est estimée au moyen de sa "profondeur" dans WordNet dans l'arbre conceptuel induite par la relation "est-un". Notons que la mesure de spécificité d'un mot est absolue, puisqu'elle est estimée par sa profondeur dans l'ontologie. Mais, elle n'est pas complètement décorrélée de l'idée d'*idf*, puisqu'un mot spécifique est souvent moins fréquent qu'un mot plus général, tout au moins pour de très grands corpus constitués d'une grande variété de textes. Cependant, ce nombre peut être qualitativement différent d'une mesure *idf* sur des corpus limités de textes spécialisés. Nous mesurons la spécificité d'un concept-terme par :

$$S_t = \#is\_a\ (t, noeud\_racine)$$

S_t est estimé par le nombre de relations *is-a* qui sépare le concept *t* du *nœud_racine* de l'ontologie.

3.2.3.5 **Taille du cluster**

La taille d'un cluster C_k, notée par $T(C_k)$ est estimée par le nombre de concepts qui le composent :

$$T(C_k) = \#t \ / \ t \in C_k$$

3.2.3.6 **Poids du cluster**

Afin de caractériser les clusters les uns vis-à-vis des autres, nous proposons de les pondérer. Plusieurs pondérations sont possibles, nous pouvons simplement choisir le poids d'un cluster en fonction de nombre de concepts qui le composent, ou même en fonction de types des relations, ou en fonction de l'importance de ses concepts. C'est cette dernière option que nous avons choisie. Le poids d'un cluster est égal à la somme des fréquences des concepts qui le composent. Formellement, le poids $P(C_k)$ d'un cluster C_k est défini par :

$$P(C_k) = \sum_{t \in C_k} f_{t,d}$$

Avec $f_{t,d}$ est la fréquence du concept *t* dans le document *d* (comme définie avant) pour chaque $t \in C_k$.

3.3 Conclusion

Dans ce chapitre, nous avons présenté une approche pour la représentation de document sous forme des clusters conceptuels ainsi que des facteurs permettant de caractériser les clusters et leurs concepts. Dans cette représentation, les relations sémantiques entre les concepts d'un document ont été utilisées pour composer les groupes de concepts, chaque groupe exprime une idée conceptuelle représentée dans un document. Ces notions seront utilisées dans les chapitres suivants : D'une part, pour proposer une sémantique d'un document, une sorte de résumé du document. D'autre part, pour sélectionner les meilleurs documents en réponse à une requête en recherche d'information.

Chapitre 4 Vue sémantique d'un texte

4.1 Introduction

Dans ce chapitre nous présentons une méthode pour l'extraction d'une vue sémantique du contenu du texte, indépendamment de toute requête. Cette vue, multi-niveaux, est censée donner un apperçu des sujets traités dans le document. Chaque niveau donne un certain détail sur le contenu du document, partant de niveau le plus général vers le plus spécifique. L'approche proposée s'appuit sur les clusters de concepts definis dans le chapitre précédent. Tout d'abord, les mots importants et significatifs sont identifiés à partir des clusters, et sont affectés aux différents niveaux de détail de la vue. Ensuite, quelques phrases contenant un maximum de ces mots sont extraites du texte pour représenter un résumé de son contenu. Cette méthode ainsi que, son intérêt est testée dans des expérimentations.

Notre objectif derrière cette notion de vue sémantique, est d'extraire les éléments (mots clés) significatifs d'un document pour permettre sa compréhension, ou encore avoir une idée de son contenu sans forcément le lire. Notre vue sémantique n'est pas un résumé du document comportant des phrases cohérentes, mais juste des « paquets » (clusters) de concepts qui donnent une idée générale du sujet traité dans le document. Cette idée peut être donnée à différents niveaux de détail du plus général vers le plus spécifique. L'approche que nous proposons dans cette section consiste à donner une vue sémantique du texte sur plusieurs niveaux. Chaque niveau est composé d'une liste des mots et donne plus de détails sur le contenu.

Ce chapitre est structuré comme suit, nous donnons tout d'abord quelques définitions sur les ensembles flous, plus précisément nous décrivons les fonctions floues que nous utilisons dans notre approche. Nous décrivons ensuite notre approche d'extraction d'une vue sémantique hiérarchique. Nous présentons les expérimentations réalisées pour mesurer l'intérêt de cette vue.

4.2 Les ensembles flous

Dans notre proposition, nous représentons notre document par des clusters sémantiques flous. Nous présentons dans cette section le concept de base des sous-ensembles flous.

Un sous ensemble flou *A* (brièvement appelé ensemble flou) d'un ensemble de référence *E* est formellement défini par une fonction d'appartenance μ_A de E dans l'intervalle des nombres réels [0,1] (degré d'appartenance qui est l'extension de la fonction caractéristique d'un sous-ensemble classique) [Zadeh, 1965][Zadeh, 1971][Kaufmann, 1973].

Pour un sous-ensemble flou *A* d'un référentiel *E* on donne les définitions suivantes :

Noyau N(A) = $\{x \,/\, \mu_A(x) = 1\}$ Les éléments «vraiment» dans A.
Support S(A) = $\{x \,/\, \mu_A(x) \neq 0\}$ Ceux qui y sont à des degrés divers.

Pour un ensemble classique *A*, noyau et support sont confondus avec *A*, et sa fonction caractéristique μ n'admet que *0* ou *1* pour valeurs.

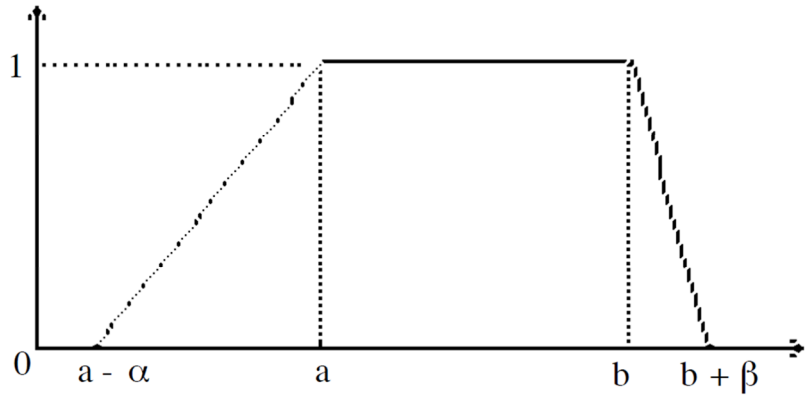


Figure 12: Ensemble flou trapézoïdal

Prenant un exemple le plus simple des ensembles flous constitué d'une représentation trapézoïde comme dans la Figure 12 qui est décrit par sa fonction d'appartenance μ_A dans un intervalle couramment utilisé « $\mathbb{R}$ ». Pour tous $x \in A$, $\mu_A(x)$ est définie par :

$$\mu_A(x) = \begin{cases} 0 \; si \; x \leq a - \alpha \; ou \; b + \beta \leq x, (x \; hors \; du \; support \; de \; A) \\ 1 \; si \; a \leq x \leq b, (x \; dans \; le \; noyau \; de \; A) \\ 1 + (x - a)/\propto \;\; si \; a - \alpha < x < a \\ (x - b - \beta)/(2b + \beta) \; si \; b < x < b + \beta \end{cases}$$

Avec la notation (a, b, ∝, β) sont des paramètres souvent utilisés dans les applications informatiques.

Dans notre proposition nous utilisons des fonctions d'appartenance dans un intervalle fermé [0, s], avec *s* est une constante. Ces fonctions sont nommées par μ_{petit}, μ_{grand}, μ_{moyen}, $\mu_{assez-grand}$, $\mu_{moyennement-grand}$ représentées suivant la Figure 13. Par exemple pour les fonctions $\mu_{petit}(x)$ *et* $\mu_{grand}(x)$, elles sont définies de la façon suivante :

$$\mu_{petit}(x) = \begin{cases} 0 \;\; s/2 < x \leq s \; (x \; hors \; du \; support \; de \; A) \\ 1 - (2 * x)/s \;\; si \;\; 0 \;\; \leq x \leq s/2 \end{cases}$$

$$\mu_{grand}(x) = \begin{cases} (2 * x)/s \;\; - 1 \; si \; s/2 < x \leq s \\ 0 \;\; si \;\; 0 \;\; \leq x \leq s/2 \; (x \; hors \; du \; support \; de \; A) \end{cases}$$

Figure 13: Partition floue des valeurs de centralité

Le principe de calcul de la fonction d'appartenance $\mu(x)$ consiste à résoudre l'équation d'une droite de la forme $\mu(x) = ax + b$ en connaissant auparavant les valeurs de x lorsque $\mu(x) = 0$ et en même temps la valeur de $\mu(x)$ lorsque $x = 0$. En général, le calcul d'une fonction d'appartenance se fait suivant un graphe comme nous l'avons déjà défini pour $\mu_{petit}(x)$ et $\mu_{grand}(x)$, nous pouvons définir $\mu_{assez-grand}(x)$ de la Figure 14.

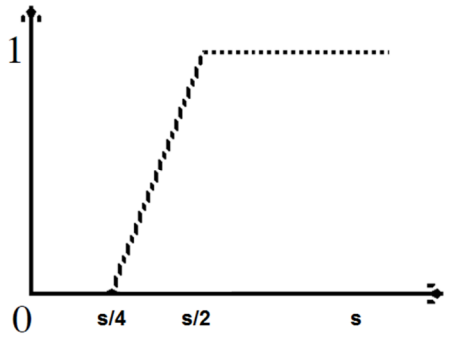


Figure 14: Exemple d'ensemble flou assez-grand

4.3 Extraction d'une vue sémantique d'un texte

Intuitivement, les mots les plus intéressants dans un texte (ici 'intéressant' veut dire capable de donner des informations sur le contenu d'un texte) sont ceux qui sont : i) fréquents, on s'appuit sur l'hypothèse classique en RI, les redondances d'un mot determinent son importance ii) en relation d'un certain type de relation avec plusieurs mots dans le texte, on s'appuit ici sur l'idée que la co-occurrence entre les mots determinent le(s) sujet(s) du document. Les termes que nous qualifions « intéressants » sont sélectionnés selon 2 hypothèses, nous utilisons tout d'abord les clusters de mots, puis dans la seconde hypothèse les mots extraits sont ceux qui ont la plus grande fréquence, la plus grande centralité ou un grand poid *tf*idf*.

L'approche que nous proposons repose sur 2 étapes. Nous commençons par représenter la méthode de sélection progressive des termes. Ensuite, nous représentons les fonctions utilisées dans la selection des mots à chaque niveau. Enfin, un algorithme détaillé est décrit.

4.3.1 Sélection progressive des termes

Le point de départ de l'approche ce sont les clusters de concepts définis dans le chapitre précédent. Pour chaque document *d*, on identifie donc l'ensemble de clusters qui le compose $\{C_1,\ldots, C_k\}$, ainsi que leur poids $P(C_k)$ (la fréquence cumulée de leurs mots, i.e. $P(C_k) = \Sigma_{w \in c_k} tf_w$). Chaque mot *w* d'un cluster est lui-même associé à trois (ou quatre) éléments d'information (ou facteurs): sa $f_{w,d}$, sa spécificité s_w, sa centralité $c_{w,d}$, (et éventuellement son idf_w). L'idée est de fournir une image du contenu d'un texte sous la forme d'un ensemble de sous-ensembles flous des mots significatifs de chaque cluster ayant un poids *important*.

L'algorithme que nous proposons est itératif, il consiste à chaque itération : i) de calculer le poids $P(C_k)$ de tous les clusters présents dans notre document, ii) d'identifier les clusters qui ont le plus fort poid (les clusters qui ont la fréquence cumulée $P(C_k)$ la plus haute (en termes de mots non encore retenus)), iii) d'extraire à partir de chaque cluster identifié, un ensemble de mots qui le représente en appliquant une certaine fonction de logique floue *hcfs(w)* que nous détaillons plus loin dans la section suivante, iv) d'éliminer des clusters les termes sélectionnés, ce qui nous permet de recalculer un nouveau poids du cluster lors de passage à la nouvelle itération (le poids d'un cluster sera mis à jour à chaque itération de la procédure).

Formellement, à chaque itération *j*, nous sélectionnons les clusters représentatifs désignés par *rep (j)* :

$$rep\ (j) = \{C_k / \ C_k\ \in clust(d)\ et\ |P(C_k) - \max\big(P(C)\big)| \leq \theta_{sc}\ \forall\ C\ \in\ clust(d)\ \}$$

Avec *clust(d)* est l'ensemble de clusters dans un document traité *d*, θ_{sc} est un seuil déterminé par les expérimentations. Nous employons « $|P(C_k) - \max\big(P(C)\big)| \leq \theta_{sc}$ » qui signifie que les clusters sélectionnés ne sont pas uniquement ceux qui ont le maximum de poids « $\max\big(P(C)\big)$ », mais en plus ceux qui sont proches de « $\max\big(P(C)\big)$ » un certain seuil θ_{sc}. Et max(P(C)) représente le poids maximal de tous les clusters.

Ensuite, à partir de chaque cluster $C_k \in rep\ (j)$, on extrait un ensemble de mots les plus significatifs qui maximise une fonction $h_{cfs}(w)$ que nous détaillons dans la section suivante. Nous désignons par $r_j(C_k)$, les termes représentatifs du cluster C_k au niveau *j* :

$$r_j(C_k) = \{w\ /\ w\ \in sel_j(C_k)\ \}\ \text{avec :}$$

$$sel_j(C_k) = \{w\ /\ w\ \in\ C_k - \cup_{n=1,j-1}\ r_n(C_k)\ et\ w\ maximise\ h_{cfs}(w)\}$$

Avec $w\ maximise\ h_{cfs}(w)$ sera détaillée dans la section suivante dans le cadre d'une démarche détaillée de sélection des mots à partir d'un cluster. Nous désignons par « $C_k - \cup_{n=1,j-1}\ r_n(C_k)$ » le cluster C_k ayant retiré les mots qui sont selectionnés aux différents niveaux avant d'arriver au niveau *j*.

La procédure de sélection est ensuite itérée sur les mots restants, en ajoutant progressivement de nouvelles couches de sous-ensembles de mots, de moins en moins significatifs, à partir des clusters qui sont progressivement sélectionnés. Au début, seuls les mots les plus significatifs des clusters les plus "lourds" sont choisis, puis davantage de clusters sont considérés, et plus de mots dans chaque cluster. Lorsque l'algorithme s'arrête de s'exécuter, un document sera représenté à différents niveaux de détails. Chaque niveau *j* est représenté par un ensemble de sous-ensembles de mots comme suit : $niveau\ j = \{r_j(C_k)\ /\ C_k\ \in rep\ (j)\ \}$. Plus on descend à un niveau, moins les termes présents à ce niveau seront représentatifs pour le document.

Le résultat peut être considéré comme un ensemble d'ensembles flous, où chaque ensemble flou correspond à l'ensemble de mots qui sont plus ou moins représentatifs (au sens de la fonction de sélection) pour un cluster donné. En effet, chaque itération fournit un ensemble (éventuellement vide) qui peut être considéré comme une section de l'ensemble flou des termes représentatifs de chaque cluster. Notons que chaque sous-ensemble flou n'est pas nécessairement normalisé. En effet, le noyau de cette structure est constitué seulement de représentants du/des cluster(s) ayant la fréquence cumulée la plus haute. En fait, nous ne sommes intéressés ici qu'à la structure stratifiée obtenue par itération de la procédure, sans avoir besoin de lui associer des degrés.

4.3.2 Selection des mots à partir des clusters

En effet, pour chaque cluster *C* sélectionné à un niveau *j*, nous faisons sortir les mots qui le représentent le mieux. Ces mots représentatifs maximisent une fonction $h_{cfs}(w)$. Nous selectionnons les mots si :

$$\big|h_{cfs}(w_i) - max\ h_{cfs}(w)\big| \leq \theta_e\ \forall\ w \in C$$

Avec θ_e un seuil de selection choisie par expérimentations.

Plus précisément, tout mot *w* appartenant à un cluster est caractérisé par les critères (ou facteurs): fréquence $f_{w,d}$, centralité $c_{w,d}$, spécificité s_w, et rareté dans la collection idf_w (dans notre cas *idf* est estimée par $\log(N/n_w)$ avec *N* est le nombre total des documents dans la collection et n_w est le nombre de documents de la collection contenant *w*). La fonction $h_{cfs}(w)$ est calculée en fonction de ces critères. Différentes fonctions ou combinaison de ces facteurs sont possibles en considérant tous les facteurs en même temps ou seulement un sous ensemble de facteurs qui pourraient être eux-mêmes combinés à d'autres facteurs, par exemple les mots qui sont les plus centraux et les plus spécifiques, ou qui ont un *tf.idf* suffisant, etc. Afin de mieux répondre à cette question, nous proposons de coder ces fonctions par des ensembles et des fonctions d'agrégation floues $\mu_{petit}(x)$, $\mu_{grand}(x)$ et $\mu_{moyen}(x)$ (les partitions floues comme décrites dans la section 4.2) avec *x* correspond aux différents facteurs utilisés.

Clairement, la fonction de sélection $h_{cfs}(w)$ prend la forme générale suivante :

$$h_{cfs}(w) = f_1\left(\mu_y(x_0), f_2\left(\mu_{y1}(x_1), \mu_{y2}(x_2)\right)\right)$$

Avec f_1, f_2 sont des fonctions telles que *min()*, *max()*. *y*, y_1 et y_2 représentent le type de la fonction floue utilisée (grand, petit, moyen, grand, assezgrand). x_0, x_1 et x_2 sont les facteurs de séléctionnement (*f, c, s, idf* ou *tf*idf*). μ est la fonction d'agrégation ayant un intervalle d'appartenance $[0, \max(x_i)]$ (i= 0, 1 ou 2 suivant le facteur employé) et des résultats dans l'intervalle des nombres réels [0,1].

Déterminer l'exacte balance entre les critères est une question d'expérimentations. Différentes fonctions d'agrégation ont été testées, dans ces fonctions nous employons les fonctions *min* et *max*. La fonction *min* est employée pour signifier que le mot selectionné satisfasse tous les critères qui la composent en même temps (exemple si j'ai besoin de selectionner les termes qui sont centraux et spécifiques on utilise $min(\mu_S, \mu_C)$). D'autre part, la fonction *max* est employée si on a besoin de sélectionner les mots qui satisfont au moins un des deux critères (par exemple la centralité ou la spécifité $max(\mu_S, \mu_C)$). De plus, dans les expérimentations, nous avons testé différentes fonctions d'agregation μ (μ_{petit}, $\mu_{assezpetit}$, μ_{grand}, $\mu_{assezgrand}$, $etc.$) pour donner plus (ou moins) de valeur à un critère donné. Prenons l'exemple pour un document *d* ayant le maximum de centralité égal à *x*, la fonction d'agrégation $\mu_{assezgrand}$ est définie suivant la Figure 12 de la section 4.2 en affectant à β zéro, à $a-\propto$ la valeur zéro, à *b* la valeur *x* et à *a* la valeur *x/2*. L'emploi de la fonction $\mu_{C-assezgrand}$ signifie que nous donnons plus de poids pour les mots qui ont une faible centralité et l'emploi de la fonction $\mu_{c-grand}$ signifie que nous annulons le poids des mots qui ont une faible centralité, ainsi nous affectons un poids uniquement pour les mots ayant une grande centralité. Dans les expérimentations, des tests effectués en combinant différents facteurs et en employant différentes fonctions d'agrégation, les fonctions h_{cfs} qui ont donné les meilleurs résultats [Bouidghaghen et al, 2008] sont :

$$h_{cfs} = max\left(\mu_{TF.IDF-assezgrand}, min\left(\mu_{S-grand}, \mu_{C-grand}\right)\right)$$

$$h_{cfs} = min\left(\mu_{TF-assezgrand}, max\left(\mu_{S-grand}, \mu_{C-grand}\right)\right)$$

A partir de la première fonction nous pouvons noter que les mots choisis sont ceux qui ont centralité et une spécifité grande ou qui ont un tf*idf assez grande. La deuxième fonction signifie que les mots choisis sont ceux qui ont une fréquence assez-grande et qui sont en même temps soit centrales, soit spécifiques.

4.3.3 Algorithme général

L'agorithme ci-dessous décrit l'alogorithme général qui décrit toute l'approche utilisée pour l'extraction des clusters et la vue sémantique

Début de l'algorithme :

(1) Extraction des noms avec une estimation de la fréquence et la spécificité

1.1 extraire les noms en utilisant TreeTagger.
1.2 Elimination des mots vides présents dans les stoplistes
1.3 Pour chaque nom w calculons $f_{w,d}$ et idf_w
1.4 Pour chaque nom w obtenir ses sens possibles en utilisant WordNet
1.5 Appliquer une méthode de désambigüisation pour choisir le sens le plus représentatif d'un mot dans le texte
1.6 Pour chaque mot w désambigüisé, calculer sa spécificité s_w qui est estimée par le nombre des relations est-un qui le sépare de root-node de WordNet. A partir de maintenant, un mot est en fait un nom désambigüisé.

(2) Construction des clusters et estimation de la centralité

2.1 Pour chaque mot *w*, construire son champ lexical CL(w) qui est défini comme un ensemble : $CL(w) = S(w) \cup I(w) \cup I^{-1}(w) \cup P(w) \cup P^{-1}(w) \cup D(w) \cup R(w)$, de WordNet. Avec $S(w)$ est un ensemble de concepts WordNet reliés à w par la relation S (pareille pour $I(w)$, $I^{-1}(w)$, $P(w)$, $P^{-1}(w)$, $D(w)$ et $R(w)$
2.2 Pour chaque mot *w*, on initialise sa centralité à 0, $c_{w,d}=0$
2.3 Pour chaque pair w_i, w_j :
Si $(CL(w_i) \cap CL(w_j)) \neq \emptyset$ alors :
$c_{wi}= c_{wi}+1$; $c_{wj}=c_{wj}+1$; mettre w_i et w_j dans le même cluster

(3) Sélection des termes représentatifs

3.1 Fixer le nombre d'itération *J* et les degrés de tolérance θ_{sc} et θ_e
3.2 Pour chaque itération $j \leq J$ repéter :
3.2.1 Pour chaque cluster C_k calculons son poids $P(C_k) = \Sigma_{w \in Ck} f(w)$
3.2.2 Calculer max $c_{w,d}$, max $f_{w,d}$, max s_w et max idf_w à partir de tous les clusters (dans l'ordre de définir la taille des domaines de définition des ensembles flous)
3.2.3 Définir les fonctions d'appartenance des ensembles flous 'moyen', 'grand', 'assez-grand' et 'tres-grand' pour chaque domaine [0, *max* c_w], [0, *max* $tf_{w,d}$], [0, *max* s_w] et [0, *max* idf_w]
3.2.4 Pour chaque mot *w* calculer $h_{cfs}(w)$ pour la fonction h_{csf} choisit.
3.2.5 Trouver le cluster qui a le poids maximal : *max P(C)*
3.2.6 Sélectionner les clusters C_k qui vérifient : $|P(C_k) - max\ P(C)| \leq \theta_{sc}$
3.2.7 Pour chaque cluster C_k obtenu dans 3.2.6 répéter :
3.2.7.1 Trouver la valeur maximal de max h_{cfs} de $h_{cfs}(w)$
3.2.7.2 Pour chaque $w_i \in C_k$ qui vérifie $|h_{cfs}(w_i) - max\ h_{cfs}| \leq \theta_e$, ajoutez le à $r_j(C_k)$
3.2.7.3 Retourner $r_j(C_k)$
3.2.7.4 Mettre à jour le cluster en cours en enlevant les mots sélectionnés, i.e. $C_k = C_k - r_j(C_k)$
3.2.7.5 mettre à jour le poids de cluster en cours en soustrayant le poids des mots sélectionnés, i.e. $P(C_k) = P(C_k) - P(r_j(C_k))$; avec $P(r_j(C_k)) = \Sigma_{w \in r_j(C_k)} tf(w)$
3.2.8 Si tous les $r_j(C_k) = \emptyset$ (i.e. il n'y a pas de mots sélectionnés à cette itération), arrêter la procédure

Fin de l'algorithme

4.4 Expérimentation et résultats

Toute évaluation en recherche d'information passe par la définition d'un protocole clair, allant de la tâche, aux métriques en passant par la collection de tests et le jugement de pertinence. La tâche que nous souhaitons évaluer est assez claire, on souhaite mesurer si les termes extraits représentent bien le contenu d'un document. Evidemment nous nous faisons de résumé de documents. De ce fait, les protocoles proposés dans ce cadre ne sont pas utilisables,

c'est pourquoi nous avons évalué cette approche selon un protocole que nous définissons dans les sections suivantes. Nous avons en fait considéré deux protocoles (ou cadre) pour l'expérimentation. Le premier évalue directement les termes extraits des documents. La seconde consiste à évaluer les phrases extraites contenant un maximum de mots significatifs.

Dans la suite de cette section, pour chaque protocole d'évaluation proposé. Nous commençons par une description des données de tests que nous utilisons. Nous décrivons ensuite un protocole d'évaluation. Nous présentons le déroulement des expérimentations. Nous finissons par une analyse des résultats obtenus.

4.4.1 Evaluation de l'extraction des termes significatifs

4.4.1.1 Collection de documents

La méthode proposée a été évaluée sur les différentes sections du document "internet" issu de Wikipedia (http://en.wikipedia.org/wiki/Internet).

Figure 15: Liste des sections internet

La Figure 15 présente la table de matière et des sections du document "internet". Pour construire notre collection, nous considérons chaque section de ce document ayant un titre comme un document. Les sections choisis sont les suivantes (1 Terminology, 2.1 Création, 2.2 Growth, 2.3 Univesity students appreciation and contributions, 3.1 Internet protocole, 3.2 Internet structure, 3.3 ICANN, 3.4 Language, 3.6 The internet viewed on mobile devices, 4.1 Email, 4.2 The world wide web, 4.3 Remote access, 4.4 Collaboration, 4.5 File sharing, 4.6 Streaming media, 4.7 Internet telephony (VoIP), 6 Internet access, 7.1 Political organization and censorship, 7.2 Leisure activities, 8 Complex architecture, 9 Marketing, 10 The terms "internet" and "internet") . Au total, nous avons 22 documents à traiter (les sections 3.5 et 5 ne sont pas considérées à cause de leurs longueurs qui ne dépassent pas les 2 lignes

considérées très courtes). En plus des 22 documents, la méthode a été appliquée sur le document entier constitué de l'ensemble de ces sections et sous sections.

4.4.1.2 **Jugement de pertinence**

La meilleure méthode pour évaluer ce type d'approches est de mener une évaluation guidée par des utilisateurs. On fait donc intervenir de vrais utilisateurs pour juger si les mots sélectionnés représentent le contenu d'un document ou pas. Ce type d'évaluation n'est pas une tâche facile en plus de temps qu'il consomme. Il s'ajoute que, sans de vrais utilisateurs, c'est difficile de construire des jugements pertinents. Afin de mettre en place cette évaluation, nous avons étudié les jugements des utilisateurs en supposant, tout simplement, que le titre d'une section ou d'un document est le meilleur représentant du contenu de ce document. Donc, l'évaluation consiste à vérifier si la représentation sémantique du document reflète le titre ou non.

Les utilisateurs doivent donc juger manuellement pour chaque document (section d'un document) si son titre correspond aux mots retirés aux premiers niveaux de notre approche. Les jugements effectués suivent le principe suivant : le titre est supposé bien identifié si les mots qui le composent apparaissent dans la liste des termes extraits.

4.4.1.3 **Déroulement des expérimentations :**

Dans les expérimentations, chaque document de la collection (inclus le document entier sur « internet ») a été traité en suivant l'algorithme cité dans la section 4.3.3. Dans l'indexation, les *concepts-termes* identifiés correspondant à des multi-termes sont désambiguïsés suivant la méthode de Baziz proposée dans 2.2.6.1. Ensuite les clusters sont construits et les différents facteurs sont calculés : *la centralité*, *la fréquence* et *la spécialité* de la même manière présentée dans la section 3.2.

Afin d'évaluer l'impact des fonctions d'agrégation, nous avons comparé deux catégories de fonctions, une première basée sur *tf.idf* et une seconde catégorie basée sur les facteurs que nous avons définis. Plus précisément, la selection de terme par *tf.idf* compte à calculer pour chaque terme du document sa fréquence notée par *tf*, ainsi *idf* calculé manuellement en utilisant google. En fait, pour chaque terme recherché sur google on obtient le nombre de documents contenant ce terme puis on estime son idf par rapport à l'ensemble des documents de google. Concernant la seconde catégorie, nous avons pris différentes fonctions d'agrégation (min, max, …) en combinant de différentes manières nos facteurs. Nous montrons seulement les résultats de celles qui offrent les meilleures performances.

Le Tableau 1 montre les résultats obtenus, la première colonne présente le titre des documents (ou sous-section du document entier), le document entier est considéré à la fin du tableau. Dans la seconde colonne, la fonction de pondération *tf.idf* est employée pour donner un poids aux mots du document, les mots tels qu'ils sont affichés apparaissent par ordre croissant de leurs poids. Dans les 4 dernières colonnes, différentes fonctions d'agrégation sont proposées, ainsi leurs résultats sont présentés comme suit : les chiffres « 1 :,2 :,3 : … » signifient les différents niveaux de représentation d'un document. A chaque niveau, nous désignons par r_i ($i=1..n$, avec n le nombre de clusters totals identifiés dans un document) les différents clusters sélectionnés, ainsi que les mots retenus à partir de chaque cluster. Par exemple, dans la 4ème colonne, le document « 2.1 Création » est représenté sur 3 itération désignée par (1 :,2 :,3 :), dans la première itération le cluster nommé r_1 a été sélectionné, à l'intérieur de ce cluster, le terme « internet » à été retiré, etc. Dans les différentes colonnes représentant les résultats, nous affichons les résultats pour au moin le second niveau sémantique obtenu par la méthode et pour un maximum de 25 mots. Les mots correspondant

aux titres du document sont marqués en gras. La dernière ligne du tableau représente le nombre de titres des documents trouvés par rapport au nombre total des documents de la collection, cette valeur est estimée en termes de pourcentage.

Document title	***tf.idf***	**Min($\mu_{s\text{-}grand}$, $\mu_{c\text{-}grand}$)**	**Max($\mu_{tf\text{-}grandsuffisamment}$, min($\mu_{s\text{-}grand}$,$\mu_{C\text{-}grand}$))**	**Max(min($\mu_{tf\text{-}grandsuffisamment}$, $\mu_{idf\text{-}grandsuffisamment}$), min($\mu_{s\text{-}grand}$, $\mu_{c\text{-}grand}$))**	**Max($\mu_{tf\text{-}assez\_grand}$, min($\mu_{s\text{-}assez\_grand}$, $\mu_{c\text{-}grand}$))**
2.1 Creation	Radar, demonstration, production, agency, satellite, ethernet, interest, Australia, summer, ease, collection, us, construction, david, service.	1: r_1={nsf, post_office, national_science_foundation}, 2: r_1={informa -tion_processing, program, e-mail}, 3: r_1={federal, air_force, darpa, information_tech -nology}.	1: r_1={internet}, 2: r_1={protocol, tcp/ip}, 3: r_1={year, system}.	1: r_1={internet}, 2: r_1={protocol, tcp/ip},	1:{}.
2.2 Growth	Word, face reference, interest, version, Protocol, year Company, majority, period, administration, internet, Switzerland.	1 : r_1={world_ wide_ web, internet}, 2: r_2={1990s}, 2: r_3={company}, 2: r_1={computer_ network}, 3: r_2={decade, year} 3: r_1={protocol}.	1: r_1={internet}, 2: r_2={1990s, decade}, 2: r_4={network}, 2: r_1={world_wide_ web}, 2: r_5={**growth**}.	1: r_1={internet}, 2: r_2={1990s, decade}, 2: r_4={network}, 2: r_1:{world_wide_web}, 2: r_5={**growth**}.	1: r_1={internet}, 2: r_1={computer_ network}, 2: r_5={**growth**}, 3: r_4={network}.
3.1 Internet protocols	Space, lan, version, task_force, world, traffic, service, capabi-lity, hardware, protocol, software, computer_system request, link, internet.	1: r_1={ip}, 2: r_1={web_site, interconnection}, 3: r_1={computer, software_system, software, engineering}.	1: r_1={**internet**}, 2: r_1={**protocol**, software, system}.	1: r_1={**internet**}, 2: r_1={**protocol**}.	1: r_1={**internet**}, 2: r_1={**protocol**}.
3.4 Language	World, asia, user, alphabet, role, capability, origin, year, Caribbean, north_america, display, internet, technology, communication, lingua_franca.	1: r_1={computer}, 2: r_2={character}, 2: r_1={technology, alphabet, display}, 3: r_3={world_ wide_web, internet}.	1: r_2={character}, 1: r_1={**language**}, 2: r_3={internet}, 2: r_1={communication, topic}.	1: r_1 = {**language**}, 2: r_2={character}, 2: r_3={internet}, 2: r_1={communication, topic}.	1: r_1={**language**}, 2: r_3={internet}, 3: r_2={character}, 3: r_1={communi-cation, topic}.
4.1 E-mail	Party, recipient, message, sender, employee, job, monitoring, timememo, internet, personnel, mail, control, mailing, system.	1: r_1={letter}, 2: r_2={party, personnel}, 2: r_1={text}.	1: r_1={letter}, 1: r_2={party, organization}, 1: r_3:{recipient}, 1: r_4={machine}, 1: r_5={control}, 1: r_6={employee}, 1: r_7:={network}, 1: r_8={job}, 1: r_9={monitoring}, 1: r_{10}={memo}, 1: r_{11}={internet}, 1: r_{12}={message, topic, content, information}, 1: r_{13}={creation}, 1: r_{14}={sender}, 1: r_{15}={concept}, 1: r_{16}={today, time}, 1: r_{17}={system}, 1: r_{18}={**e-mail**}, 1: r_{19}={detail}, 2: r_1={mailing, mail, text}, 2: r_2={personnel}.	1: r_1={letter}, 1: r_2={party, organization}, 1: r_3:{recipient}, 1: r_4={machine}, 1: r_5={control}, 1: r_6={employee}, 1: r_7:={network}, 1: r_8={job}, 1: r_9={monitoring}, 1: r_{10}={memo}, 1: r_{11}={internet}, 1: r_{12}={message, topic, content, information}, 1: r_{13}={creation}, 1: r_{14}={sender}, 1: r_{15}={concept}, 1: r_{16}={today, time}, 1: r_{17}={system}, 1: r_{18}={**e-mail**}, 1: r_{19}={detail}, 2: r_1={mailing}, 2: r_2={personnel}.	1: r_{18}={**e-mail**}, 2: r_{11}={internet}, 3: r_2={party, organization}, 3: r_7={network}, 3: r_{17}={system}.
4.7 Internet telephony	Feature, fee, voice, extension,	1: r_1={ip}, 1: r_2={electronics},	1: r_1={ip}, 1: r_4={**internet**},	1: r_1={ip}, 1: r_4={**internet**},	1: r_4={**internet**}, 1: r_5={voice},

(VoIP)	world, traffic, service, reliability, direction, protocol, power, year, **internet**, power_failure, electronics.	1: r_3={cable, **telephone**, phone, modem}, 2: r_1={protocol}, 3: r_1={interopera-bility}.	1: r_2={system}, 1: r_3={**telephone**}, 1: r_5={voice}, 2: r_1={interopera-bility}, 2: r_6={call}, 2: r_3={phone}, 2: r_7={year}, 2: r_5={communi-cation}, 2: r_8={gaming}, 2: r_9={emergency}, 2: r_{10}={provider}.	1: r_2={system}, 1: r_3={**telephone**}, 1: r_5={voice}, 2: r_1={interopera-bility}, 2: r_6={call}, 2: r_3={phone}, 2: r_7={year}, 2: r_5={communi-cation}, 2: r_8={gaming}, 2: r_9={emergency}, 2: r_{10}={provider}.	2: r_6={call}, 2: r_3={**telephone**}, 2: r_{10}= {provider}, 3: r_1={ip}, 3: r_2={system}, 3: r_5={communi-cation}.
Internet (The whole document)	Demonstration, republic, day, production, lan, standing, movie, satellite, party, side, task_force, collection, image, asia, service, encyclopedia, channel, relaxation, remote, distinction, alphabet, june.	1: r_1={file_server, desktop}, 2: r_1={ship}, 3: r_1={business, information_proces s-ing, ip, text_file, program, email, e-mail, television}.	1: r_1={**internet**}, 2: r_1={protocol, network}, 3: r_1={people, user, computer, software, information, world_wide_web, system, term}, 4: r_1={ip, communication, video, provider}, 5: r_1 ={year, computer_network, e-mail, company, work, layer, file}.	1: r_1={**internet**}, 2: r_1={network}, 3: r_1={protocol}, 4: r_1={system}, 5: r_1={user}, 6: r_1={software}, 7: r_1={computer, information, world_wide_web}.	1: r_1={**internet**}, 2: r_1={network}, 3: r_1={protocol}, 4: r_1={system}, 5: r_1={user}.
Nombre de titres trouvés/ nombre de document de la collection	7/23 (**30%**)	9/23 (**40%**)	17/23 (**74%**)	17/23 (**74%**)	16/23 (**70%**)

Tableau 1: Extrait de résultats obtenus avec quelques fonctions d'agrégation sur le document "internet" et sur les ses sous-sections

4.4.1.4 **Analyse des résultats et discussion**

A partir des résultats obtenus, nous pouvons tirer différents résultats qui justifient l'intérêt de notre méthode :

Au niveau du document entier : L'application de la méthode de clustering sur le document entier a généré un cluster principal regroupant la majorité des mots du document, ainsi que de petits clusters de quelques mots et un certain nombre de singletons. Le poids de cluster principal domine les poids des autres clusters. Ainsi, l'apparition d'un cluster dominant a deux effets principaux. D'une part, le document est composé d'une idée principale représentée par un seul cluster. D'autre part, à chaque itération de l'algorithme d'extraction des mots significatifs, uniquement ce cluster est selectionné comme on peut le constater dans le Tableau 1.

Nous constatons que l'apparition d'un grand cluster n'a pas influencé la performance des résultats. En effet, dans les trois fonctions d'aggregation que nous utilisons, les résultats présentent les différents sujets traités dans le document. Ainsi le titre du document apparaît au premier niveau de ces résultats, pour les 3 fonctions d'agrégation composées d'une combinaison de la centralité et la spécificité avec la fréquence comparant à la fonction tf*idf où le titre n'apparaît pas.

Au niveau des fonctions d'agrégations : Les résultats obtenus par les fonctions d'agrégation des différents critères *tf*, *idf*, *c* et *s* sont clairement meilleurs que ceux obtenus par la fonction *tf.idf*. En fait, dans les résultats du Tableau 1 nous avons présenté en gras les titres des sections qui apparaissent dans les résultats. Par exemple, pour la section 3.1

"internet protocole", les mots "internet" et "protocole" sont retournés dans les premières itérations de nos fonctions d'agrégation, tandis que, dans les résultats retournés par *tf.idf,* les mots "internet" et "protocole" apparaissent après une liste des termes non significatifs.

Pour chaque fonction d'agrégation, nous commentons la dernière ligne du Tableau 1 estimant le nombre des titres trouvés par rapport au nombre total des documents. Pour la fonction $\mathbf{max}(\mu_{\text{tf-assezgrand}}, \mathbf{min}(\mu_{\text{s-assezgrand}}, \mu_{\text{c-grand}}))$, on constate qu'elle retourne exactement les titres des documents dès la première itération (et uniquement les titres). Cette fonction sélectionne uniquement les termes qui sont très spécifiques et très centraux ou très fréquents. En appliquant cette fonction sur nos 23 documents, elle retourne les titres des documents dans les premières itérations pour 16 /23 documents, ce qui est équivalent à 70% des titres de documents qui ont été retrouvés.

Ensuite, en utilisant les 2 fonctions $\mathbf{max}(\mu_{\text{tf-grandsuffisament}}, \mathbf{min}(\mu_{\text{s-grand}}, \mu_{\text{c-grand}}))$ et $\mathbf{max}(\mathbf{min}(\mu_{\text{tf-grandsuffisament}}, \mu_{\textit{idf}\text{-grandsuffisament}}), \mathbf{min}(\mu_{\text{s-grand}}, \mu_{\text{c-grand}}))$, qui correspondent respectivement aux colonnes 4 et 5 du Tableau 1, retournent presque les mêmes résultats. Il faut noter que pour 17 de nos 23 documents, les titres ont été identifiés dans les premières itérations. Nous pouvons constater que l'introduction de critère *idf* ne retourne pas des meilleurs avantages.

Par contre, en appliquant la fonction $\mathbf{min}(\mu_{\text{s-grand}}, \mu_{\text{c-grand}})$ en intégrant uniquement les critères *c* et *s* sur nos 23 documents comme présentée dans la colonne 3 du Tableau 1, uniquement 9 titres ont été identifiés.

Cas particuliers : Nous examinons en plus le cas des mauvais résultats, comme le cas pour le document nommé "creation". La représentation de ce document ne reflète pas ni son contenu, ni son titre. Ce problème est lié à la nature du document qui contient beaucoup d'expressions chronologiques (des années) et des noms propres qui ne sont pas pris en compte par notre approche, parce qu'ils ne sont pas considérés par l'ontologie que nous avons utilisée.

A partir de cette expérimentation, nous pouvons conclure que la méthode d'extraction d'une vue sémantique d'un texte appliquée sur l'identification des titres ramène de très bons résultats. Nous avons également montré que la prise en compte de la fréquence en combinaison avec la centralité et la spécifité rend des meilleurs résultats comparés à ceux obtenus par tf*idf. En fait, en utilisant la fonction tf*idf, uniquement 30% des titres ont été trouvé comparant à 70% obtenus par les fonctions d'agrégation que nous avons proposées.

4.4.2 Extraction de phrases significatives

Dans le but de montrer l'efficacité de notre méthode d'extraction d'une vue sémantique d'un texte, nous l'avons utilisée dans une seconde expérimentation dans le but d'extraire les phrases significatives d'un texte. La technique d'extraction de telles phrases s'appuit entièrement sur la méthode d'extraction des termes significatifs décrits tout au long de ce chapitre. Les phrases extraites par la méhtode sont celles contenant un maximum des termes identifiés aux premiers niveaux de la méthode d'extraction d'une vue sémantique. Dans cette expérimentation, nous comparons les phrases que nous extrayons automatiquement par notre approche à celles séléctionnées manuellement par des utilisateurs.

Plus précisément, l'approche d'extraction des phrases significatives consiste à extraire à partir d'un texte les phrases supposées représentatives suivant plusieurs critères. Idéalement, un ensemble de phrases sera d'autant plus convenable pour donner une bonne idée du contenu d'un texte que i) un nombre maximum de mots parmi les plus représentatifs y figurant, ii) les mots issus d'un maximum de clusters y apparaîssant, iii) l'ensemble des phrases sélectionnées

restera court (la longueur des phrases étant mesurée par le nombre de mots constituant la phrase).

On pourrait envisager pour cette expérimentation, de représenter en termes de critères flous, l'objectif à atteindre. De plus, une procédure heuristique plus simple est expérimentée qui sélectionne d'abord, la phrase contenant « *la proportion de mots significatifs la plus grande*». A titre indicatif, toutes les phrases (ou sous-titres) dont la proportion de ses termes significatifs sont supérieures à 1/4 sont données.

Nous présentons dans ce qui suit la collection des documents que nous utilisons dans les expérimentations ainsi que le protocole que nous utilisons pour sélectionner les phrases pertinents. Nous représentons ensuite le déroulement des expérimentations. Nous finissons par une analyse des résultats obtenus.

4.4.2.1 **Collection de documents**

Pour cette expérimentation, nous avons collecté 20 articles de différentes tailles extraits des différents journaux portant sur « le nauffrage de l'Erika » qui a eu lieu en 1999 et ses conséquences. Les liens vers les documents composant notre collection de tests sont listés dans le Tableau 2. Dans la première colonne du Tableau 2, nous attribuons un numéro pour chaque article. Dans la seconde colonne nous donnons le lien vers l'article. Ces documents traitent l'accident selon différents points de vue (quelques documents s'intéressent sur l'effet environnmental, d'autres mettent le focus sur les effets juridiques…).

Doc n°	**Lien vers le Document**
1	http://english.aljazeera.net/news/europe/2008/01/200852512481572296.html
2	http://www.birdlife.org/news/news/2008/01/Erika_judgement.html
3	http://news.bbc.co.uk/1/hi/world/europe/592378.stm
4	http://www.planetark.org/dailynewsstory.cfm/newsid/12876/newsDate/18-Oct-2001/story.htm
5	http://www.ifremer.fr/docelec/notice/2004/notice403-EN.htm
6	http://www.foei.org/en/publications/annual-report/2007/what-we-achieved-in-2007/member-group-victories/europe/france-huge-symbolic-victory-in-erika-oil-spill-trial
7	http://cybertheses.francophonie.org/index.php/record/view/22342
8	http://www.efluxmedia.com/news_French_Oil_Company_Total_Found_Guilty_for_Erika_Oil_Disaster_12935.html
9	http://www.guardian.co.uk/environment/2008/jan/17/oilspills.pollution
10	http://www.imo.org/Environment/mainframe.asp?topic_id=231
11	http://www.abc.net.au/4corners/archives/2000b_Monday25September2000.htm
12	http://www.charlestannock.com/180100.asp
13	http://www.marinelink.com/Story/Company-Knew-Tanker-was-Risk-Before-Erika-Disaster-205959.html
14	http://afp.google.com/article/ALeqM5hhlV25s200BELhe7RWZ32X13bCyg
15	http://www.alertnet.org/thenews/newsdesk/L11816302.html
16	http://www.mg.co.za/article/2008-01-17-30-000-tonne-oil-disaster-costs-total-200m
17	http://celticcountries.com/webmagazine/environment/erika-oil-spill-brittany-legal-precedent-maritime-pollution/
18	http://www.chemgapedia.de/vsengine/vlu/vsc/en/ch/16/uc/vlus/erikaoilspill.vlu/Page/vsc/en/ch/16/uc/oilspill/casestudies/erika/erikaenvironment.vscml.html
19	http://www.iht.com/articles/2007/06/06/business/total.php
20	http://www.reuters.com/article/environmentNews/idUSL0427183620070604

Tableau 2: Liens vers les documents de la collection

4.4.2.2 **Jugement de la pertinence**

Afin de juger la pertinence de nos résultats, nous avons fait intervenir des utilisateurs afin d'extraire manuellement à partir de chaque document de la collection ses thèmes (ses sujets). Le thème désigne une ou plusieurs phrases qui correspondent soit à des phrases extraites directement du texte, soit à des phrases reformulées représentant une idée significative du document.

Les phrases extraites automatiquement sont alors comparées aux thèmes extraits manuellement. Nous considérons qu'un thème extrait manuellement a été identifié par une ou plusieures phrases extraites par notre méhtode dans 2 cas : premier cas, si la phrase extraite par notre méthode correspond exactement à un thème extrait manuellement. Dans le second cas, si la phrase extraite par la méthode représente l'idée présentée par un thème manuel.

Pour chaque document nous présentons le nombre des thèmes trouvés par rapport au nombre total des thèmes du document. Une moyenne des thèmes trouvés est calculée pour l'ensemble des documents de la collection.

4.4.2.3 **Déroulement de l'experimentation**

Nous procédons dans les expérimentations par une représentation multi-niveau du chaque document en appliquant l'algorithme décrit dans la section 4.3.3. Les *concepts-termes* représentatifs des documents extraits aux différents niveaux correspondent à des concepts WordNet multi-termes. Les concepts sont désambiguïsés en utilisant la méthode de Baziz listée dans la section 2.2.6.1 pour avoir un document représenté par une liste des *concepts-termes* désambiguïsés. Les relations entre les *concepts-termes* sont identifiées, ainsi les clusters sont construits et les valeurs de *la centralité*, *la fréquence* et *la spécialité* sont calculées.

Nous menons notre expérimentation sur une seule fonction d'agrégation, celle qui a fourni les meilleurs résultats dans l'expérimentation précédente, soit $\mathbf{max(\mu_{tf\text{-}grandsuffisament}, min(\mu_{s\text{-}grand}, \mu_{c\text{-}grand}))}$. Ensuite nous appliquons notre règle d'extraction des phrases significatives (les phrases du texte contenant ¼ des termes représentatifs sont sélectionnées manuellement).

Nous présentons dans le Tableau 3 un échantillon des résultats sur quelques documents de la collection. Dans la première colonne, nous présentons le numéro du document traité. Dans la seconde colonne, nous montrons la liste des phrases extraites pour chaque document par notre méthode (les termes extraits aux différents niveaux de la méthode d'extraction d'une vue sémantique sont représentés en annexe 3145, dans le Tableau 19). La troisième colonne liste les thèmes identifiés par les utilisateurs.

La compatibilité entre les thèmes de la méhtode et ceux des utilisateurs est identifiée par un même nombre d'étoiles dans les deux colonnes. Les thèmes des utilisateurs qui n'ont pas été identifiés par les phrases extraites par notre méhtode sont marqués par des tirets et réciproquement.

Doc n°	**Phrases extraits du texte**	**Thèmes générale du document extrait manuellement**
1	* Total guilty over Erika oil spill. ** French oil giant Total has been ordered to pay millions of dollars in damages after being found responsible for the 1999 sinking of the tanker Erika, one of France's worst environmental disasters.	* French oil giant Total responsible for the 1999 sinking of the tanker Erika, ** French court had held the charterer of a tanker responsible for pollution caused through shipwreck.

	** Eleven others, including the ship's captain, were found not guilty. *** 'Severe warning' - The defendants could face hundreds of millions of dollars in further damages after the court said environmental organisations could sue them over the ecological impact of the disaster. * Toxic fuel ** The case finally came to trial in February 2007. - 'Murky world'	*** Plaintiffs accused the company of negligence in hiring the ship and of acting too slowly when the accident happened.
7	* A qualitative and quantitative assessment was conducted of this contamination in water, suspended particulate matter, sediments, and in intertidal molluscs. * (4) consistent and visible temporal decline in concentrations for water, SPM and molluscs,	*An investigation was carried out into the PAH chemical contamination resulting from the "Erika" tanker fuel spillage ** The results of this study demonstrated that heavily oil-contaminated shorelines… *** The increase in the contamination levels before and after the spill, together with the significant change in the pattern of PAH composition
10	*The Erika disaster and revised single-hull phaseout schedule ** On 12 December 1999 the 37,238-dwt tanker ** Erika broke in two in heavy seas off the coast of Brittany, France, while carrying approximately 30,000 tonnes of heavy fuel oil	* The Erika disaster and revised single-hull phaseout schedule ** circumstances of the Erica accident
15	*Total on trial over 1999 French oil disaster * MONSTER TRIAL * The French government alone is seeking 153 million euros. * Besides Total and two of its subsidiaries, the ship's Indian captain, its management company, four French maritime officials and the Italian maritime certification company RINA, which classified the ship as safe, are also on trial. ** Critics, including the environmental group Friends of the Earth, which is one of the plaintiffs in the trial, say Total took cynical risks with the ship to meet a tight contract deadline.	* Opening of the trial on the Erica disaster - Erica accident and effect on the environment - Political are among some 70 plaintiffs **Total accused of marine pollution and endangering human lives
17	* Erika oil spill in Brittany could set legal precedent for responsibility in maritime pollution ** Brittay's worst-ever environmental disaster	* Erika oil spill in Brittany could set legal precedent for responsibility in maritime pollution **Brittay's worst-ever environmental disaster - Fishing and tourism business were severely damaged
20	* Prosecutor wants Total convicted for Erika disaster ** PARIS (Reuters) - French oil giant Total should be convicted of maritime pollution for its role in the sinking of the oil tanker Erika, which provoked one of France's worst environmental disasters, prosecutors said on Monday. - The company denies the charges	* Prosecutor wants Total convicted for Erika disaster - Erica history and effects seabirds ** Total failed to conduct proper checks before chartering the ageing ship. ** Total had faced pollution and negligence charges as well as complicity in endangering human lives over the incident. * Prosecution convict six other individuals and organizations

Tableau 3: Les phrases extraites manuellement comparant à celles extraites automatiquement

4.4.2.4 **Analyse des résultats et discussion**

Le tableau 4.3 a été ensuite analysé pour mesurer le nombre des phrases « pertinents » effectivement trouvés par notre approche. Le tableau 4.4 liste les résultats obtenus pour l'ensemble des documents considérés pertinents. La première colonne liste le numéro du document, la seconde liste le nombre de termes constituant chaque document avec les mots vides, la troisième colonne présente les différents facteurs obtenus par la fonction « **$\max(\mu_{\text{tf-grandsuffisament}}, \min(\mu_{\text{s-grand}}, \mu_{\text{c-grand}}))$** ». Ainsi, nous présentons dans la colonne «**N° d'itération** » le nombre d'itération de la procédure de sélection, la colonne « **Nb des clusters sélectionnés** » présente pour chaque document les clusters qui ont participés à la sélection des termes. La colonne « **Nb des termes retournés** » présente le nombre total des termes

sélectionnés dans les différentes itérations de notre méthode. La colonne « **Thèmes trouvés / existent thèmes (% pert)** » présente le rapport entre le nombre de thèmes trouvés par notre méthode et le nombre de thèmes extraits manuellement, ce rapport est quantifié en terme de pourcentage. La dernière ligne de ce tableau présente un moyen de résultats pour les 20 documents.

Doc n°	Nb des termes	Résultat $\max(\mu_{tf\text{-}grandsuffisament}, \min(\mu_{s\text{-}grand}, \mu_{c\text{-}grand}))$			
		N° d'itération	Nb des clusters sélectionnés	Nb des termes retournés	Thèmes trouvés / existent thèmes (% pert)
1	446	5	20	51	3/3 (100%)
2	301	3	14	25	3/4 (75%)
3	473	3	12	14	2/2 (100%)
4	267	3	11	21	2/2 (100%)
5	209	3	13	19	2/2 (100%)
6	181	4	9	21	3/3 (100%)
7	342	5	12	28	2/3(66,66%)
8	345	3	21	46	2/2 (100%)
9	1099	8	8	55	5/6 (83,33%)
10	106	1	6	12	2/2 (100%)
11	209	2	6	19	2/2 (100%)
12	362	3	17	30	2/3(66,66%)
13	538	6	3	13	2/4 (50%)
14	608	4	12	20	4/5 (80%)
15	518	6	15	58	4/4 (100%)
16	496	7	28	62	3/4 (75%)
17	175	2	15	23	2/3(66,66%)
18	176	4	11	29	2/2 (100%)
19	515	5	5	17	3/4 (75%)
20	295	2	15	41	4/5 (80%)
Moyenne	380	4	12,5	30	2,7/3,25 = 83,07 %

Tableau 4: Résultats obtenus pour l'ensemble de documents

Les résultats comme présentés dans le Tableau 4 montrent une grande compatibilité entre les thèmes extraits manuellement et les phrases du document extraites automatiquement. En fait, pour la plupart des documents, on a pu identifier 100% des thèmes manuels, pour une moyenne de 83% pour tous les documents de la collection.

Notons que la moyenne des termes extraits par document en utilisant notre méthode d'extraction des thèmes est de 30 termes, ce qui est équivalent à 8% de la longueur moyenne des documents (380 termes). De plus le nombre moyen de clusters est égal à 12,5 ce qui signifie qu'il y a une varieté dans la sélection des termes à partir de plusieurs clusters, ce qui est un bon signe dans la composition des phrases extraites.

Les phrases extraites par la méthode donnent une bonne idée du texte en général, ces phrases sont beaucoup plus faciles à lire et à comprendre pour un utilisateur que des clusters de termes isolés. Ces phrases peuvent être utilisées dans différents domaines, tels que les résumés textuels ou pour représenter les phrases d'un snippet dans un moteur de recherche.

Malgré les résultats que nous avons obtenus, nous pouvons noter certaines limites qui sont reliées à WordNet. D'une part, certains mots qui sont intéressants pour représenter le contenu d'un document ne sont pas présents dans la base de WordNet, exemple les mots comme "biomarkers" et "polycyclic-aromatic-hydrocarbon", ce qui donne que ces mots ne peuvent pas être présents dans la liste finale des concepts représentatifs d'un document. D'autre part, les *concept-termes* extraits de WordNet peuvent ne pas représentés les mots ou groupes de mots qui sont réellement présents dans le document si par exemple ils ont été mal désambigüisés. Ceci se traduit par la non détection du vrai sens, ce qui nous empêche de créer de bons clusters. Exemple: dans le document 11 le nom de reporteur de BBC "Mangold" est représenté par le sens d'une plante. Dans le document 12 le mot "oil" est désambigüisé par "vegetable oil" pourtant il désigne "fuel oil".

D'autres limites ont été également constatées, la présence de clusters contenant des termes comme "year, month, week, day,…", signifie que le document contient des informations historiques. Les termes sont mals sélectionnés à partir de ces clusters, de tels termes pareils nécessitent une procédure spécifique pour les exploiter.

4.5 Conclusion

Dans ce chapitre, nous avons présenté une méthode permettant de construire une vue sémantique d'un document. Cette vue permet de comprendre de quoi parle un document sans forcément le lire. La méthode s'appuit sur un processus d'extraction des termes significatifs d'un document. Ces termes sont extraits de manière progressive en s'appuyant sur deux critères : leur fréquence et leur centralité dans le document.

Nous avons conduit deux types d'évaluation, la première a pour objectif de vérifier si la représentation sémantique de document reflète le titre du document traité. Dans cette évaluation nous avons fait intervenir des utilisateurs pour juger manuellement pour chaque document (section d'un document) si son titre correspond aux mots retirés aux premiers niveaux de notre approche. Les résultats obtenus dans cette expérimentation montrent que 70% des titres ont été identifiés comparant à 30% obtenus par tf*idf.

La seconde consiste à utiliser les termes significatifs représentant la vue semantique pour extraire les phrases significatives d'un texte. Les phrases extraites par la méhtode sont celles contenant un maximum des termes identifiés aux premiers niveaux de la méthode d'extraction d'une vue sémantique. Dans cette expérimentation, nous avons comparé les phrases que nous extrayons automatiquement par notre approche à celles séléctionnées manuellement par des utilisateurs. Pour la plupart des documents testés, on a pu identifier 100% des thèmes manuelles, pour une moyenne de 83% pour tous les documents de la collection.

Chapitre 5 Nouveaux facteurs pour la Recherche d'Information

5.1 Introduction

La majorité des modèles de RI utilisent les facteurs *tf*, *idf*, taille du document, d'une part pour mesurer la pertinence d'un document vis-à-vis d'une requête. D'autre part, ils s'appuient principalement sur la détection de traits morphologiques d'un document pour caractériser son contenu. Nous proposons pour notre part d'introduire de nouveaux facteurs permettant à notre sens, de mieux caractériser la dimension sémantique du contenu d'un document. Ces facteurs vont au-delà du simple comptage des termes, ils tentent de mesurer combien un terme donné est relié (sémantiquement) aux autres termes d'un document. Ces facteurs sont la « *centralité* », la « *fréquence conceptuelle* » et la « *spécificité* » définis dans le chapitre 3. Dans ce chapitre nous étudions l'intérêt de ces différents facteurs dans deux cadres différents. Nous proposons tout d'abord une technique de désambiguïsation de concept basée sur la centralité. Ensuite, nous introduisons ces facteurs pour mesurer la pertinence d'un document vis-à-vis d'une requête (tâche de RI ad-hoc).

Ce chapitre est organisé de la façon suivante. Nous décrivons tout d'abord notre méthode de désambiguïsation de concepts d'un document basée sur la *centralité*. Nous proposons ensuite dans la section 5.3 une fonction de pondération de termes basée sur la *centralité*, la *fréquence conceptuelle* et la *spécificité*. Les sections suivantes sont consacrées aux expérimentations. Nous avons tout d'abord décrit les expérimentations d'une façon générale. Ensuite nous avons effectué une série d'expérimentation sur deux collections de documents de TREC afin de choisir la meilleure configuration, combinaison, des paramètres que nous avons introduits pour mesurer la pertinence d'un document vis-à-vis d'une requête. Nous sommes ensuite intéressés à mesurer l'impact des relations sémantiques (synonymie, hyperonymie, …) qui sont au cœur des facteurs centralité et fréquence conceptuelle sur les performances.

5.2 Désambiguïsation basée sur la centralité

Nous avons présenté dans la section 2.2.6.1 du rapport, les différentes étapes permettant de représenter un document par un sac de concepts désambiguïsés suivant la méthode de désambiguïsation proposée par Baziz [Baziz, 2005].

Dans cette section, nous proposons une méthode de désambiguïsation des termes basée sur la centralité. Cette méthode est composée de plusieurs étapes. La première étape consiste à utiliser la méthode d'extraction des termes (*simples ou composés*) à partir d'un document *d* comme nous l'avons déjà présentée dans la section 2.2.6.1. A l'issue de cette étape, le document *d* est représenté par un ensemble de termes R_d :

$$R_d = \{t_i \,/\, i = 1, k(d)\}$$

Avec t_i est un terme simple ou composé (multi-terme) présent dans le document *d*. *k(d)* est le nombre des termes dans le document *d*. Pour les mêmes raisons évoquées dans la section 3.2.1 on ne grande dans R_d que les noms.

Chaque terme t_i de R_d peut avoir *n* entrées (ou *concept-terme*) possibles dans WordNet, donc plusieurs sens possibles dont chacun est noté par C_j^{ti} pour *j=1,… n*. Chaque sens

appartient à un synset de WordNet. Nous notons l'ensemble de ces sens possibles pour un terme t_i, par S_{ti} :

$$S_{ti} = \{ C_1^{ti}, C_2^{ti}, \dots, C_n^{ti} \}$$

Le terme t_i a donc $|S_{ti}|=n$ sens, chaque sens est rattaché à un synset de WordNet, c'est-à-dire, à chaque terme t_i correspond n entrées dans WordNet.

La deuxième étape de la méthode consiste à calculer pour chaque *concept-terme* C_i^{ti} de S_{ti} un poids qui correspond à sa centralité dans le document *d,* notée par "$c_{ci,d}$" de la même manière introduite dans la section 3.2.3.

La méthode de désambiguïsation que nous proposons considère que le concept-terme ayant la plus grande centralité est le meilleur représentant du sens du terme dans le document. Formellement, le meilleur concept attaché à un terme est donné par :

$$C_{ti}^{*} = argmax\left(centralité\left(C_j^{ti}\right) / \; C_j^{ti} \in S_{ti}\right)$$

A l'issue de cette étape, nous obtenons une représentation réduite du document *d* par un ensemble des concepts désambiguïsés, noté RS_d :

$$RS_d = \{ C_{ti}^{*} / \; \forall \; t_i \; \in R_d \; \}$$

Pour évaluer l'intérêt de notre méthode nous la comparons avec deux autres méthodes de désambiguïsation. La première méthode, considère pour chaque terme le premier sens du concept qui lui correspond dans WordNet. La seconde est celle proposée par Baziz [Baziz, 2005].

5.3 La fonction c×f×s basé sur *c*, *f* et *S* pour la RI

Les modèles de RI que nous avons décrits dans le chapitre 1 se basent comme nous l'avons signalé sur des facteurs *tf*, *idf*, *dl* pour pondérer un document vis-à-vis une requête, dont la fonction RSV est souvent une combinaison de ces facteurs, elle prend souvent la forme suivante:

$$\mathrm{RSV}(\mathrm{q}, \mathrm{d}) = f(tf, idf, n, N, dl)$$

Où *tf* représente la fréquence d'apparition de chaque terme de la requête *q* dans le document *d*. *idf* est la fréquence inverse calculée pour les termes de la requête et basée sur le nombre de document de la collection contenant le terme noté *n* et le nombre total des documents dans la collection de test noté *N*. *dl* représente la longueur de *d* qui correspond au nombre des termes qui le compose.

Nous rappelons que nous utilisons concept-terme pour désigner un sens d'un terme appartenant à un concept (synset) WordNet comme nous l'avons présenté dans la section 2.3.1.

Il est bien connu en RI que des facteurs tels que *tf*, *idf* et la longueur du document sont efficaces pour les approches de recherche d'information basées sur les termes simples. Ces facteurs ne sont pas suffisants pour mesurer l'intérêt d'un concept (ou d'une phrase). Les facteurs que nous proposons peuvent être considérés comme de nouvelles sources d'évidence qui peuvent être rajoutées à n'importe quel schéma de pondération comme celui proposé par Baziz [Baziz, 2005], Robertson [Robertson et al, 1981], Salton [Salton et al, 1988], Jin [Rong et al, 2005] and Croft [Korvetz et Croft, 1992]. Ces facteurs sont la centralité, la fréquence conceptuelle et la spécificité. Notre intuition derrière ces facteurs quant à leur intérêt en RI est la suivante :

Centralité : D'une part, nous pensons que ce facteur permet de mieux sélectionner les termes centraux d'un document. En fait, un terme qui a le plus grand nombre de relations avec les concepts de document est plus significatif pour représenter son contenu. D'autre part, ce facteur permet d'estimer un poids pour n'importe quel terme par rapport à n'importe quel document même si ce dernier n'existe pas dans le document. Ce facteur est intéressant lorsque nous estimons le poids pour des termes d'une requête qui n'existent pas dans un document.

La spécifité : Ce facteur est identique à *idf* permet de donner plus de valeur à des termes spécifiques. Plus un concept est spécifique (profond) dans l'arbre conceptuel de l'ontologie utilisée, plus son sens est précis.

La centralité, la fréquence et la spécificité peuvent être intégrés dans des modèles de RI existants, de même ils peuvent être utilisés seuls ou combinés avec d'autres facteurs. Nous avons proposé une fonction d'appariement RSV(q,d) d'une requête *q* vis-à-vis un document *d* permettant de les combiner seuls :

$$\mathrm{RSV(q,d)} = \sum_{t \in q} \left(\left(C_{t,d} \right)^{\alpha} * \left(S_t \right)^{\beta} * \left(1 + F_{t,d} \right)^{\gamma} \right)$$

Avec *d* et *q* sont représentés sous forme de concepts-termes issus de WordNet. α, β *et* γ sont des paramètres à optimiser. $C_{t,d}$ est la centralité du concept-terme *t* de la requête *q* dans *d*. S_t est la spécificité du concept-terme *t* dans WordNet. $F_{t,d}$ est la fréquence du concept *t* dans le document. Dans les expérimentations, nous varions les valeurs de α, β *et* γ pour évaluer l'impact de différents facteurs.

5.4 Expérimentations et résultats

Nous décrivons dans cette section les différentes expérimentations que nous avons menées pour mettre à l'épreuve nos deux propositions : la technique de désambiguïstion et les facteurs centralité, fréquence conceptuelle et spécificité. Nous avons vu tout au long du chapitre 3 que les facteurs *centralité* et *fréquence conceptuelle* peuvent être calculés en fonction de différents types de relations et ils peuvent être combinés de différentes façon quand ils sont utilisés en RI. Nous allons tenter de varier tous ces paramètres pour comprendre leur impact réel.

Dans la suite de cette section nous décrivons tout d'abord les collections de test que nous utilisons. Nous montrons ensuite le protocole d'évaluation de pertinence que nous utilisons. Nous continuons par une description de différentes étapes que nous utilisons dans les expérimentations. Nous définissons une notation pour désigner les différentes combinaisons des paramètres utilisés lors des expérimentations.

5.4.1 Collections de test TREC

Dans les expérimentations, nous utilisons deux collections de TREC[13] distribuées par NIST « TREC1 et TREC7 ». Ces deux collections sont essentiellement composées de 3 parties : les documents de la collection, les topiques et les jugements de pertinence.

Les documents de la collection TREC1 sont partagés sur deux disques (D1 & D2) d'une taille d'environ 1GB chacun et provenant des ressources différentes. Le disque D1 est composé des ressources suivantes :

- WSJ -- Wall Street Journal (1986, 1987, 1988, 1989)
- AP -- AP Newswire (1989)
- ZIFF -- Information from Computer Select disks

13 http://www.nist.gov/

- (Ziff-Davis Publishing)
- FR -- Federal Register (1989)
- DOE -- Short abstracts from Department of Energy

Et le disque D2 est composé des ressources suivantes :

- WSJ -- Wall Street Journal (1990, 1991, 1992)
- AP -- AP Newswire (1988)
- ZIFF -- Information from Computer Select disks
- (Ziff-Davis Publishing)
- FR -- Federal Register (1988)

Les documents de la collection TREC7 sont partagés sur deux disques (D4 & D5) provenant des ressources différentes, dont le disque D4 est composé des ressources suivantes :

- Financial Times Limited (1991, 1992, 1993, 1994)
- The Congressional Record of the 103rd Congress (1993)
- The Federal Register (1994)

Et le disque D5 est composé des ressources suivantes :

- Foreign Broadcast Information Service (1996)
- The Los Angeles Times (1989, 1990).

Les cinquantes topiques de TREC1 (resp. TREC7) sont numérotés 51-100 (resp. 351-400) ont été conçus pour imiter le besoin réel d'un utilisateur. Ils sont rédigés par de vrais utilisateurs de recherche d'information, exemple de requêtes :

```
<top>
<num> Number: 351
<title> Falkland petroleum exploration
<desc> Description:
What information is available on petroleum
exploration in the South Atlantic near the Falkland
Islands?
<narr> Narrative:
Any document discussing petroleum exploration in the
South Atlantic near the Falkland Islands is
considered relevant.  Documents discussing petroleum
exploration in continental South America are not
relevant.
</top>
```

Dans nos expérimentations nous ne considérons que le texte de la balise <title> pour constuire la requête. Le but étant d'être proche des requêtes de vrais utilisateurs qui sont souvent composées de quelques termes.

5.4.2 Protocole d'évaluation

5.4.2.1 **Baseline**

Les documents sont premièrement indexés en utilisant une indexation des termes classiques en sélectionnant les termes simples qui apparaissent dans les documents. Ensuite, une méthode de lemmatisation est appliquée en utilisant l'algorithme de porter [Porter, 1980] et enfin, les mots vides sont éliminés en utilisant une stoplist standard [Salton et McGill, 1983]. Un poids est affecté à chaque terme en suivant la formule de pondération BM25 de Robertson et Walker [Robertson et Walker, 1994a]. Le même processus est appliqué à la

requête. Chaque requête est soumise au système Mercure [Boughanem et al, 2003], le système retourne les top 1000 documents. Ces résultats sont considérés comme la baseline.

Par les expérimentations que nous avons menées dans le cadre de ces travaux, nous nous sommes limités à appliquer nos approches à un sous-ensemble de documents qui comportent au moins un terme de la requête. Pour des raisons de simplicité et de réduction des temps de traitement, nous prenons les 100 premiers documents renvoyés par la baseline.

5.4.2.2 **Mesures d'évaluations**

Les résultats sont évalués selon la précision Px ou MAP. La précision Px à un point x est une proportion évaluant le nombre des documents pertinents retirés à partir des tops x documents. En plus, pour une meilleure évaluation des résultats obtenus, surtout lorsque nous comparons deux méthodes, nous effectuons un test de signifiance statistique de type t-test. Le t-test consiste à assigner une valeur de confidence (p-valeur) pour l'hypothèse nulle. L'hypothèse nulle typique c'est qu'il n'y a pas de différence entre les deux systèmes. Lorsque la p-valeur est faible, typiquement si p-valeur <0.05, l'hypothèse nulle est rejetée (ce qui signifie que les résultats ne sont pas obtenus par chance).

5.4.3 Les étapes d'expérimentations

Dans chaque expérience que nous menons, on suit les étapes suivantes :

Extraction des concepts : La première étape consiste à extraire à partir de chaque document l'ensemble des termes susceptibles de représenter des concepts de l'ontologie. Pour cela, avant d'éliminer tous les mots vides du document (ceux de stoplistes, en français, les mots "de", "un", "les", etc. sont les plus fréquents, en anglais, ce sont "of", "the", etc.), nous détectons les termes composés par des mots uniques ou des groupes de mots. Ces termes peuvent correspondre à différentes entrées (ou nœuds) dans l'ontologie. Dans nos expériences, pour la simplicité nous ne gardons que les noms, nous nous basons sur l'ontologie de WordNet.

Désambiguïsation document : La deuxième étape consiste à choisir pour chaque terme identifié dans un document le meilleur sens (concept-terme) WordNet qui lui correspond dans le document. Dans cette étape, nous avons utilisé trois méthodes de désambiguïsation. La première méthode est la plus simple, dans cette méthode il n'y a pas de detection de sens, nous représentons un terme par le concept-terme WordNet qui correspond au premier sens. La deuxième méthode est celle proposée par Baziz comme décrite dans la section 2.2.6.1. Dans la troisième méthode nous employons notre méthode de désambiguïsation proposée dans la section 5.2.

Désambiguïsation requête : La troisième étape consiste à identifier pour chaque terme de la requête un concept-terme (ou ensemble de concept-terme) qui lui correspond suivant la méthode de désambiguïsation utilisée. Ensuite pour chaque concept requête nous calculons la valeur de ces différents facteurs ($C_{t,d}$, $f_{t,d}$, $Cf_{t,d}$, S_t , $P(C_k)$) par rapport à chaque document de la baseline comme présenté dans la section 3.2.3 du chapitre 3. Nous avons utilisé dans cette étape trois méthodes de représentation des concepts-termes. La première méthode nommée « *premier sens du concept* » consiste à représenter chaque terme identifié dans la requête par le concept-terme qui correspond au premier sens. La deuxième méthode nommée « *tous les sens possibles* » consiste à représenter les termes de la requête par tous les sens (concepts-termes) possibles. La troisième méthode nommée « *désambiguïsation par la centralité* » consiste à représenter chaque requête par une liste de concepts-termes convenable à un document interrogé. Dans cette dernière méthode, nous choisissons parmi les concepts-termes

possible de représenter un sens d'un terme le concept-terme ayant la plus grande centralité dans le document intérrogé.

Appariement requête-document : La quatrième étape consiste à calculer un score de pertinence d'une requête vis-à-vis un document. Là aussi, en considérant différentes variations de cette fonction selon les facteurs ci-dessus.

Evaluation des résultats : La dernière étape consiste à évaluer l'efficacité des résultats obtenus en calculant les valeurs du rappel et de la précision. Ensuite pour une meilleure évaluation du système, une comparaison de la précision obtenue est effectuée par rapport à d'autres modèles.

5.4.4 Configurations utilisées

Chaque expérience menée correspond à une configuration particulière constituée des différentes méthodes utilisées dans différentes étapes (*Extraction des concepts*, *Désambiguïsation document*, *Désambiguïsation requête*) comme présenté dans le Tableau 5.

Pour ne pas réexpliquer à chaque expérience les différentes méthodes utilisées, nous notons une configuration par *combinaison-paramètre(x,y,z).* Le paramètre *x* de cette fonction *(resp. y et z)* représente la méthode utilisée pour l'extraction des concepts *(resp. méthode utilisée pour la désambiguïsation du document et la méthode utilisée pour la désambiguïsation des requêtes)* où:

- *x* correspond à la méthode d'extraction des concepts. Elle prend la valeur « conc_mt » pour signifier (concept multi-terme) si nous identifions à partir du document les groupes de mots (ou multi-termes) qui correspondent à des concepts de WordNet. Elle prend la valeur « conc_s » pour signifier (concept simple) si nous identifions à partir d'un document les termes simples qui correspondent à des concepts WordNet.

- *y* correspond à la méthode de désambigüisation des concepts du document. Avec *y* prend la valeur « doc_sens1 » correspond à une méthode de représentation de concept par son premier sens, la valeur « doc_dB » correspond à la méthode de désambiguïsation proposée par Baziz [Baziz, 2005] et la valeur « doc_dC » correspond à notre propre méthode de désambiguïsation par la centralité.

- z correspond à la méthode que nous utilisons pour représenter les concepts d'une requête. Avec *z* prend la valeur « req_sens1 » pour indiquer que nous utilisons la méthode qui correspond au *premier sens du concept*, la valeur « req_tous-sens » pour indiquer que nous utilisons la méthode *tous les sens possibles* qui correpond à l'extension des concepts de la requête par les synonymes et la valeur « req_dC » pour indiquer que nous utilisons la méthode *désambiguïsation par la centralité* qui correspond à la désambiguïsation basée sur la centralité.

Extraction des concepts basée sur:	**Désambigüisation des concepts du document**	**Désambigüisation des concepts d'une requête**
conc_mt : multi-termes	*doc_sens1* : Le premier sens du concept	*req_sens1* : Le premier sens du concept
conc_s : termes simples	*doc_dB* : désambiguïsation par la méthode de Baziz	*req_tous-sens* : Tous les sens possibles
	doc_dC : désambigüisation en se basant sur la centralité	*req_dC* : Désambiguïsation par la centralité

Tableau 5: Différentes variantes utilisés dans les expérimentations

5.5 Meilleures valeurs de *α, β et γ* de notre fonction d'appariement

Dans le but d'évaluer l'impact de chacun des facteurs « la centralité » et « la spécificité » ainsi que de leur combinaison avec « la fréquence », nous suivons une liste des expérimentations menées sur la collection TREC1 qui consistent à trouver pour notre fonction d'appariement les meilleures valeurs de *α, β et γ*. Dans ces expérimentation, nous considérons la configuration suivante : *combinaison-paramètre (*conc_mt *, doc_dB, req_dC).* Puisqu'au départ nous considérons que l'emploi des multi-termes est intéressant dans l'indexation, ensuite Baziz [Baziz, 2005] a montré que sa méthode de désambiguïsation pour les concepts du document est efficace. Pour la méthode de désambiguïsation des concepts de la requête nous avons considéré que l'emploi de la centralité est la meilleure méthode comme nous le prouvons plus tard dans les expérimentations. Nous avons fait varier *α, β et γ* de la fonction d'appariement qui suit :

$$\mathrm{RSV}(\mathrm{q},\mathrm{d}) = \sum_{t \in q} \left(\left(C_{t,d}\right)^{\alpha} * \left(S_t\right)^{\beta} * \left(1 + F_{t,d}\right)^{\gamma} \right)$$

Dans cette fonction, afin d'éviter la multiplication par zéro, nous ajoutons un à la fréquence. Les résultats obtenus peuvent prendre différentes valeurs. Nous listons dans le Tableau 6 les résultats obtenus en variant les différentes valeurs de *α, β et γ*. Nous simplifions la représentation des différents facteurs dans le Tableau 6 avec *C* désigne $C_{t,d}$, *F* désigne $F_{t,d}$ et *S* désigne S_t.

A partir des premiers résultats nous constatons que l'utilisation de la centralité est plus efficace que la spécificité et la fréquence. Cependant, le fait de rajouter *F* à la centralité améliore la pertinence, ensuite *S* améliore plus sur le top des documents. Les meilleurs résultats sont obtenus par la combinaison des trois paramètres comme présenté par la colonne 7 (respectivement 8) pour α = 1 (respectivement 2), bien que pour α=2 on obtient plus de gain au niveau de précision.Nous gardons pour les expérimentations suivantes la formule de la colonne 8 avec $\alpha = 2$, $\beta = 1$ *et* $\gamma = 1$.

RSV	∑ C	∑ S	∑ F	∑C*S	∑C*(1+F)	∑C²*F	∑C*(1+F)*S	∑C²*(1+F)*S
P5	0,4833	0,3167	0,4508	0,4917	0,4917	0.4833	0,5042	0,500
P10	0,4625	0,3312	0,4083	0,4500	0,4917	0.4917	0,4792	0,4917
P15	0,4444	0,3556	0,4014	0,4403	0,4667	0.4694	0,4681	0,4694
P20	0,4344	0,3583	0,3969	0,4198	0,4469	0.4552	0,4469	0,4562
P30	0,4174	0,3528	0,3931	0,4125	0,4257	0.4375	0,4264	0,4375
P100	0,3633	0,3633	0,3633	0,3633	0,3633	0.3633	0,3633	0,3633
MAP	0.0929	0.0799	0,0848	0.0928	0.0930	0.0943	0.0925	0.0937

Tableau 6: Impact des différents paramètres

5.6 Optimiser les paramètres de « combinaison-paramètre(x,y,z) »

Dans cette section nous souhaitons étudier l'impact des différentes méthodes pouvant influencer la recherche d'information (*méthodes d'extractions des concepts, méthode de désambiguïsation des concepts document, méthode de désambiguïsation des concepts requête*).

La procédure que nous suivons dans cette partie consiste à comparer l'impact de chacun des paramètres de la fonction *combinaison-paramètre(x,y,z)* sur notre fonction. Nous menons nos expérimentations sur la collection TREC1. La première section consiste à comparer les différentes méthodes de désambiguïsation et de choisir la meilleure pour désambiguïser les concepts d'un document et d'une requête. Dans la seconde section nous montrons que la prise en compte des mutli-termes pour la représentation du contenu d'un document n'est pas très significative. Dans les sections suivantes nous comparons nos résultats à différentes fonctions connues telle que BM25 du modèle de probabilité OKAPI [Robertson et Walker, 1994a], et à une méthode d'indexation conceptuelle proposée par Baziz [Baziz, 2005]. De plus nous proposons une nouvelle fonction pour combiner notre fonction d'appariement avec BM25.

5.6.1 Evaluation des techniques de désambiguïsation des concepts requête et document

Pour comparer les résultats obtenus par les différentes méthodes de désambiguïsation de concept d'un document et d'une requête, dans ces expérimentations nous varions les différentes valeurs de x, *y* et *z* de la *combinaison-paramètre(x,y,z)*. Les différentes combinaisons sont comparées à la *combinaison-paramètre (conc_MT, doc_dC, req_dC)* correspondant à l'extraction des multi-terme d'un document et à la desambiguïsation requête et document par la centralité.

Nous présentons dans le Tableau 7 les précisions obtenues par les différentes méthodes de désambiguïsation document. Chaque colonne présente une combinaison de paramètres de *combinaison-paramètre(x, y , z)* représentée par *x-y-z* (exemple la colonne libellée *conc_MT-doc_sens1-req_dC* signifie l'utilisation de la fonction *combinaison-paramètre(conc_MT, doc_sens1, req_dC)*). Les précisions obtenues par les différentes méthodes sont comparées à la precision obtenue dans la première colonne qui combine en même temps la désambiguisation requête et document par la centralité.

	conc_MT-doc_dC-req_dC	*conc_MT-doc_sens1-req_sens1 (%imp)*	*conc_MT-doc_sens1-req_dC (%imp)*	*conc_MT-doc_sens1-req_tous-sens (%imp)*	*conc_MT-doc_dB-req_sens1 (%imp)*	*conc_MT-doc_dB-req_dC (%imp)*
P5	**0,525**	0,4186 (-20,27%)	0,493 (-6,10%)	0,5023 (-4,32%)	0,4708 (-10,32%)	0,5 (-4,76%)
P10	**0,4938**	0,414 (-16,16%)	0,4651 (-5,81%)	0,4698 (-4,86%)	0,4396 (-10,98%)	0,4817 (-2,45%)
P15	**0,4667**	0,4 (-14,29%)	0,4589 (-1,67%)	0,4496 (-3,66%)	0,4153 (-11,01%)	0,4604 (-1,35%)
P20	**0,4656**	0,4012 (-13,483%)	0,45 (-3,35%)	0,4372 (-6,10%)	0,4063 (-12,74%)	0,4562 (-2,02%)
P30	**0,4472**	0,3853 (-13,84%)	0,4217 (-5,70%)	0,4225 (-5,52%)	0,3993 (-10,71%)	0,4375 (-2,17%)
P100	**0,3633**	0,3633	0,3633	0,3633	0,3633	0,3633
MAP	**0,093**	0,0735	0,0813	0,0811	0,0871	0,0915

Tableau 7: Comparison des résultats obtenus par les différentes méthodes de désambiguïsation de concept document

A partir des résultats obtenus dans le Tableau 7 nous pouvons constater que la méthode représentée dans la seconde colonne du tableau qui utilise la désambiguïsation documents par la centralité et la désambiguïsation requête par la centralité rend les meilleurs résultats avec une amélioration de la précision à tous les niveaux d'un pourcentage équivalent à 5%. Ainsi nous constatons que dans les colonnes où l'on utilise la désambiguïsation par la centralité (colonne 2,4 et 7 pour les requêtes et 2 pour les documents) on aura une augmentation au niveau de précision, comparons la méthode représentée dans la 3ème colonne qui n'utilise pas la désambiguisation par centralité ni au niveau document, ni au niveau document on a une amélioration de précision d'un pourcentage >5%. De même, comparée à la méthode de désambiguïsation proposée par Baziz, les résultats sont listés dans les colonnes 6 et 7, nous constatons également une amélioration significative >5%. Pour toutes ces raisons nous pouvons confirmer que l'utilisation de la méthode de désambiguïsation par centralité est efficace. Dans les expérimentations suivantes on garde cette méthode de désambiguïsation basée sur la centralité.

5.6.2 Prise en compte des multi-termes vs terme simple

Le processus d'identification des concepts multi-termes d'un document correspond à une tâche lourde au niveau de la programmation. Ce processus, comme nous l'avons décrit dans la section 2.2.6.1 consiste à trouver des groupes de termes du document susceptibles de représenter des concepts de WordNet. Dans le cas d'identification des concepts multi-termes, les concepts identifiés peuvent correspondre à des groupes de termes simples qui se suivent dans le document comme *pull_one's_weight* qui est composé des termes simples *pull*, *one's* et *weight*.

Dans cette section nous testons si l'emploi des multi-termes ramène une véritable amélioration au niveau de la précision. Pour ce faire, nous remplaçons le processus d'identification des concepts multi-termes par un autre processus plus simple. Nous identifions uniquement pour les termes simples extraits du document les concepts qui leur

correspondent dans WordNet, sans aller chercher si un groupe de mots simple du document correspond à un concept WordNet. Ensuite, le processus de désambiguïsation que nous employons sur ces concepts est celui de notre méthode basé sur la centralité. Ainsi, le processus de calcul de la centralité de ces concepts est celui proposé dans la section 3.2.3.

Les résultats que nous présentons dans le Tableau 8 comparent les résultats obtenus en employant les termes simples à celles utilisant les multi-termes pour notre fonction de pondération « c^2*f*s ». Dans ce tableau nous représentons dans la seconde colonne les résultats obtenus en appliquant les deux méthodes d'indexation document et requête par la centralité en considérant les concepts multi-termes *combinaison-paramètre(conc_MT, doc_dC, req_dC).* Dans la troisième colonne nous utilisons les mêmes méthodes d'indexation document et requête mais en employant les termes simples *combinaison-paramètre(conc_S, doc_dC, req_dC)*. Dans la quatrième colonne, nous représentons le pourcentage d'amélioration obtenu en employant les concepts qui correspondent à des termes simples.

Les résultats obtenus montrent que l'utilisation des termes simples amène une légère amélioration, ce qui signifie que l'emploi des concepts multi-termes n'est pas très significatif. Cela est dû à notre sens à différentes raisons. La première vient de WordNet, nous pensons qu'il existe des termes significatifs qui ne sont pas présent dans WordNet. La seconde vient de la difficulté de pondérer les concepts multi-termes, la notion de fréquence n'est probablement pas adéquate. Enfin, la tâche de détection de concepts multi-termes n'est pas une tâche facile surtout que les documents indexés ont un certain nombre d'erreurs syntaxiques, ainsi des termes qui se suivent avec une erreur syntaxique evitent la détection des concepts multi-termes. Pour ces différentes raisons, l'emploi des concepts multi-termes ne nous paraît pas utile, pour la suite nous faisons nos expérimentations avec une indexation conceptuelle basée sur les termes simples.

	combinaison-paramètre(conc_MT, doc_dC, req_dC)	*combinaison-paramètre(conc_S, doc_dC, req_dC)*	(%imp)
P5	0,5250	0,5375	2,38%
P10	0,4938	0,4917	-0,43%
P15	0,4667	0,4861	4,16%
P20	0,4656	0,4635	-0,45%
P30	0,4472	0,4417	-1,23%
P100	0,3633	0,3633	0,00%
map	0,0930	0,0973	4,62%

Tableau 8: Impact des concept multi-termes

5.6.3 Evaluation de notre modèle d'appariement avec deux approches

Nous avons ensuite comparé notre modèle d'appariement avec deux modèles de la littérature : le modèle BM25, correspondant à notre baseline, et le modèle proposé par Baziz basé sur le même principe de concepts-terme et désambiguïsation que nous utilisons dans notre approche.

5.6.3.1 Comparaison de notre modèle d'appariement à BM25

Nous comparons nos meilleurs résultats (obtenus par la configuration *combinaison-paramètre conc_S, doc_dC, req_dC)* à la baseline obtenu en utilisant la formule BM25 sur les termes simples. Ainsi, dans le but de tester l'influence de la combinaison d'une fonction à

base fréquentielle avec notre méthode, nous proposons une nouvelle fonction qui est une combinaison de la fonction BM25 avec notre fonction, définie comme suit :

$$Comb(BM25, c^2 * f * s) = \alpha \times Normal(BM25) + (1 - \alpha) \times Normal(c^2 * f * s)$$

Où α est une constante. Normal(BM25) (resp. Normal(c²*f*s)) est une valeur normalisée de BM25 (resp. c²*f*s) entre [0,1]. Avec la normalisation de scores ramène les valeurs entre la borne supérieure qui correspond au max de scores obtenus pour chaque requête selon BM25 (resp. c²*f*s) et le borne inférieur qui correspond à la min des scores obtenu pour chaque requête en appliquant BM25 (resp. c²*f*s) à l'interval [0,1].

Le Tableau 9 liste la précision pour les top-x document obtenus pour les 50 requêtes utilisées par les évaluations de la collection TREC1. Dans la deuxième colonne nous présentons les top-x précisions obtenues pour les 100 documents de la baseline. La troisième colonne représente la précision obtenue par notre fonction suivie du pourcentage d'amélioration obtenu par notre fonction comparée à la baseline. Dans la quatrième colonne, nous présentons les top-x précisions obtenues par la fonction qui combine BM25 avec notre méthode suivie du pourcentage d'amélioration par rapport à la baseline. Nous présentons dans la dernière ligne de ce tableau, la valeur de la précision moyenne calculée sur l'ensemble des requêtes ainsi la valeur t-test (Student Test) pour confirmer qu'il s'agit d'un taux significatif.

Nous constatons clairement que les résultats obtenus par notre approche sont meilleurs que ceux de la baseline. La troisième colonne du Tableau 9 liste une forte amélioration pour les top_x documents. Ainsi que la valeur t-test obtenue est <5%, ce qui confirme que l'amélioration est significative. A partir des résultats présents dans la quatrième colonne nous pouvons constater que la combinaison de BM25 avec une méthode conceptuelle ramène une amélioration comparée à la baseline et à notre fonction d'appariement. Ces améliorations sont justifiées plus tard dans la section 5.7 lors d'une étude menée sur la distribution de pertinences dans la liste des documents retournés par BM25 vs ceux retournés par notre fonction. En effet, nous montrons que les documents pertinents classés en premier dans la liste des résultats retournés par une fonction sont classés plus loin dans la liste retournée par l'autre et inversement. La valeur de t-test est <5%, confirme l'amélioration obtenue.

	BM25	∑C²*(1+F)*S (% imp)	Comb(BM25, c²*f*s) (% imp)
P5	0,4500	0,5375 (19,44%)	0,5542 (23,16%)
P10	0,4437	0,4917 (10,82%)	0,5142 (15,89%)
P15	0,4375	0,4861 (11,11%)	0,5042 (15,25%)
P20	0,4323	0,4635 (7,22%)	0,4917 (13,74%)
P30	0,4194	0,4417 (5,32%)	0,4674 (11,44%)
P100	0,3633	0,3633 (0,00%)	0,3633 (0,00%)
Map	0,0930	0,0973 (4,62%)	0,1042 (12,04%)
t-test		t-test p <5%	t-test p<5%

Tableau 9: Comparer nos résultats à celles obtenus par BM25

En comparant notre approche à BM25, on peut supposer que l'amélioration est due à l'emploi des concepts dans les requêtes et les documents. En effet BM25 utilise les mots simples alors que dans notre approche nous identifions pour chacun de ces mots un concept qui lui correspond dans une ontologie, le fait qui nous a permis de sortir des nouveaux facteurs sémantiques tels que la centralité. L'emploi des concepts améliore les résultats, nous avons alors comparé notre approche à une approche conceptuelle basée sur WordNet, celle proposée par Baziz [Baziz, 2005].

5.6.3.2 **Comparaison de notre approche à celle de Baziz [Baziz, 2005]**

Dans cette section, on compare nos résultats à ceux obtenus par une méthode conceptuelle proposée par Baziz [Baziz, 2005] pour la même collection TREC1.

Dans cette approche, les documents et les requêtes sont représentés sous forme des concepts WordNet désambiguïsés comme présenté dans la section 2.2.6.1. L'appariement document requête est mesuré de la manière suivante :

$$\mathrm{RSV}(q,d) = \sum_{t \in q} (\mathrm{C\_score}(t) + \mathrm{Classic}(t))$$

Avec classique est la mesure obtenue par la formule BM25 pour un terme, et $\mathrm{C\_score}(t)$ est un score conceptuel basé sur une mesure de similarité de Resnik [Resnik, 1999][Baziz, 2005].

	Baziz	∑C²*(1+F)*S	(%imp)
P5	0,3167	0,5375	69,72%
P10	0,3313	0,4917	48,42%
P15	0,3417	0,4861	42,26%
P20	0,3521	0,4635	31,64%
P30	0,3542	0,4417	24,70%
P100	0,3633	0,3633	0,00%
map	0,0811	0,0973	19,98%

Tableau 10: Comparaison entre la precision obtenue par nos résultats et celle obtenue par une méthode conceptuelle

Comme présenté dans le Tableau 10, on peut noter que la précision obtenue par notre modèle est largement meilleure, à tous les niveaux, que celle obtenue par le modèle de Baziz. La troisième colonne présente le pourcentage des améliorations.

5.7 Influence de la centralité et de la spécificité sur la pertinence

Dans la section 5.6.3, nous avons montré que les résultats obtenus par « BM25 » et « c²*f*s » sont différents. Nous rappelons que la fonction « c²*f*s » a été employée pour retrier les 100 premiers documents selectionnés par BM25 pour chacune des 50 requêtes de la collection TREC1. Dans cette section nous menons une étude pour vérifier si les documents pertinents classés en premier pour chacune des deux fonctions sont les mêmes.

Nous nous limitons à faire cette étude sur les 10 premiers documents selectionnés pour la requête 59. Pour chacun de ces documents nous comparons sa position dans la liste des documents selectionnés par les 2 approches.

Les résultats sont représentés par deux tableaux dont chacun représente uniquement les 7 premiers documents pertinents sélctionnés par BM25 dans le premier tableau « Tableau 11 » et par c²*f*s dans « Tableau 12 ». La première colonne de chacun des deux tableaux donne le numéro de document selectionné. Dans la colonne 2, 3 et 4, nous listons pour chaque concept-terme de la requête les valeurs des différents facteurs avec c représente la centralité, s représente la spécificité et f représente la fréquence. La colonne « Ordre BM25 » donne la position des documents selon BM25 et la colonne « Ordre c²*f*s » donne la position de ces documents selon c²*f*s.

N° doc pertinent	fatality#n#1			fatality#n#2			weather#n#1			Ordre BM25	Ordre c²*f*s
	c	s	**f**	c	s	**f**	c	s	**f**		
AP890227-0016	0	8	**1**	0	7	**1**	3	8	**7**	3	10
AP880904-0049	0	8	**1**	1	7	**1**	1	8	**6**	12	27
AP891219-0082	1	8	**2**	0	7	**2**	5	8	**5**	14	5
AP890827-0011	1	8	**1**	0	7	**1**	2	8	**3**	17	26
AP881212-0134	0	8	**2**	0	7	**2**	6	8	**6**	18	2
AP880614-0027	2	8	**2**	0	7	**2**	0	8	**2**	19	29
AP881212-0089	0	8	**1**	0	7	**1**	6	8	**6**	29	3

Tableau 11: Les dix premiers documents pertinents selectionnés selon BM25

A partir du Tableau 11 nous constatons qu'entre les 10 premiers documents de la liste selon BM25, un seul document « AP890227-0016 » pertinent a été classé troisième. Il faut aller jusqu'aux 29 documents pour trouver 7 documents pertinents.

N° doc pertinent	fatality#n#1			fatality#n#2			weather#n#1			Ordre BM25	Ordre c²*f*s
	c	s	**f**	c	s	**f**	c	s	**f**		
AP891115-0199	1	8	**1**	0	7	**1**	8	8	**10**	40	1
AP881212-0134	0	8	**2**	0	7	**2**	6	8	**6**	18	2
AP881212-0089	0	8	**1**	0	7	**1**	6	8	**6**	29	3
AP891128-0070	1	8	**2**	0	7	**2**	7	8	**4**	37	4
AP891219-0082	1	8	**2**	0	7	**2**	5	8	**5**	14	5
AP890308-0202	0	8	**1**	0	7	**1**	6	8	**4**	61	8
AP890227-0016	0	8	**1**	0	7	**1**	3	8	**7**	3	10

Tableau 12: Les dix premiers documents pertinents retirés selon BM25

A partir du Tableau 12 nous constatons que les 7 premiers documents pertinents selectionnés selon c²*f*s ont été classés dans les 10 premiers documents de la liste. Comparés à leur ancien ordre selon BM25 nous constatons qu'ils étaient placés loin dans la liste (colonne « Ordre BM25 » du tableau).

Nous constatons ainsi en comparant les deux tableaux « Tableau 11 et Tableau 12 » que nous ne récupérons pas les mêmes documents dans les 2 approches. De plus, nous remarquons que la fréquence n'est pas un facteur suffisant pour selectionner les documents pertinents, ainsi sa combinaison avec des facteurs comme la centralité et la fréquence permet de mieux les selectionner dans la liste. D'autre part, nous pouvons constater que les documents pertinents classés en premier pour une fonction à base fréquentiel, ne sont pas les mêmes classés en premier pour notre fonction à base de centralité. Notons que les documents selectionnés par c²*f*s représente une forte centralité de « weather », le fait qui a augmenté leurs poids pour qu'ils soient classés premiers dans la liste, cela montre l'intérêt de la centralité pour selectionner les documents pertinents.

5.8 Impact des relations sur la centralité et la fréquence conceptuelle

Le calcul de la *centralité* et de la *fréquence conceptuelle* d'un concept *t* dans un document *d* est basé sur les relations WordNet existantes entre les concepts de *d* noté par $R(t, t_k)\ \forall\ t_k\ \in\ d$. Avec $R(t, t_k)$ signifie qu'il existe une relation X entre *t* et t_k où X = S ∪ I ∪ I^{-1} ∪ P ∪ P^{-1} ∪ D ∪ R, avec S, I, P, D et R sont les relations de base de WordNet

(Synonymie, hyperonymie ou spécialité, hyponymie ou est-une-partie-de, dérivé et domaine) qui peuvent exister entre deux mots. La relation glossaire n'est pas traité vu son lourd traitement qui est très couteux en termes de calcul.

Le but dans cette section est d'évaluer l'impact des relations de la *centralité* et de la *fréquence conceptuelle*. Dans les expérimentations menées nous calculons la *centralité* et la *fréquence conceptuelle* en utilisant uniquement les relations directes entre les concepts (deux concepts d'un même document sont considérés reliés s'ils sont directement X-relié).

Par la suite, dans chaque expérience menée nous précisons le type des relations que nous avons utilisé. Alors nous notons : $R(S, I, P, D, R)$ pour représenter tous les types des relations utilisées entre les concepts, où chaque paramètre de R désigne un type de relation et qui sera représenté par « 1 » si la relation représentée par ce paramètre est présente et « 0 » si la relation n'existe pas. Dans toutes les expérimentations menées, nous prenons S = 1 pour désigner que nous utilisons toujours la relation de synonymie puisque nous considérons que deux termes synonymes doivent être toujours reliés.

Cette section est décomposée en deux sections. La première consiste à évaluer l'impact de différents types de relations sur le calcul de *centralité*. Dans la seconde nous étudions l'impact de différentes relations sur la *fréquence conceptuelle*.

5.8.1 Impact des relations WordNet pour représenter la centralité

Dans le but de faire une étude complète sur la centralité, nous étudions l'impact de chacune des relations WordNet sur le calcul de la centralité d'un concept dans un document. Dans les expérimentations précédentes, la centralité d'un concept *t* dans un document *d* notée par $c_{t,d}$ est estimée par le nombre de concept dans *d* qui sont X-reliés à *t*. Avec *X* représente n'importe quel type de relation existante entre concept WordNet, de même X peut être une relation indirecte entre deux concepts t_i et t_j d'un document *d* à travers un tiers concept t_k non présent dans *d* et qui est X-relié à t_i et X-relié à t_j. Dans cette section, pour mieux comprendre l'intérêt de chacune des relations WordNet dans le calcul de la centralité, nous menons des expérimentations pour étudier l'impact de chaque relation dans $R(S, I, P, D, R)$ sur le calcul de la centralité "$c_{t,d}$".

Ces expérimentations sont réalisées sur deux collections différentes issues de la collection TREC de NICT « TREC1 » et « TREC7 » Dans ces expérimentations, nous utilisons les termes simples dans l'indexation ainsi que les meilleures méthodes de désambigüisation par centralité pour les documents et les requêtes obtenus précédemment.

Nous utilisons dans les expérimentations notre fonction d'appariement « $c^2*(1+f)*s$ ». Les résultats obtenus pour chaque combinaison de relations sont présentés sous formes graphiques (Figure 16 pour la collection TREC1, ainsi, la Figure 17 pour la collection TREC7). Les valeurs de la précision pour top-10 documents sont représentées graphiquement dans les figures pour chaque combinaison des relations possibles de R. La combinaison R1(1,1,1,1,1) correspond à l'approche telle que nous avons testée précédement c.à.d. nous employons toutes les relations pas seulement en considérant les relations directes mais en plus les relations indirectes entre les concepts. Les résultats complets incluant les tableaux et les graphiques sont listés dans la partie annexe 3.2.

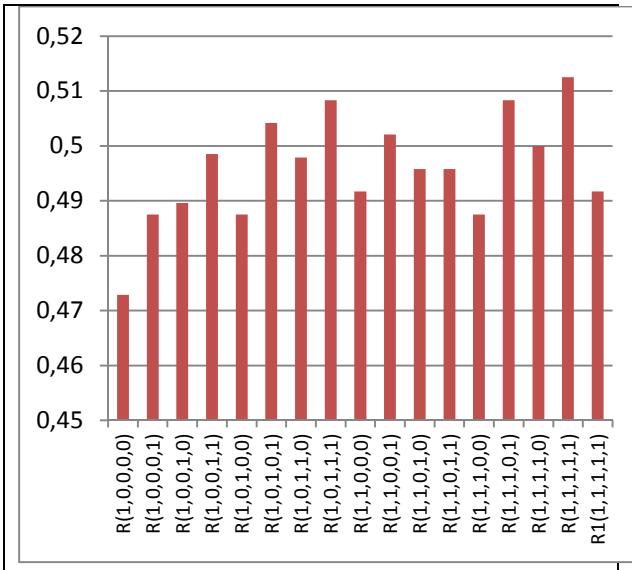


Figure 16: Comparaison graphique de la précision pour les différentes relations sur la collection TREC1

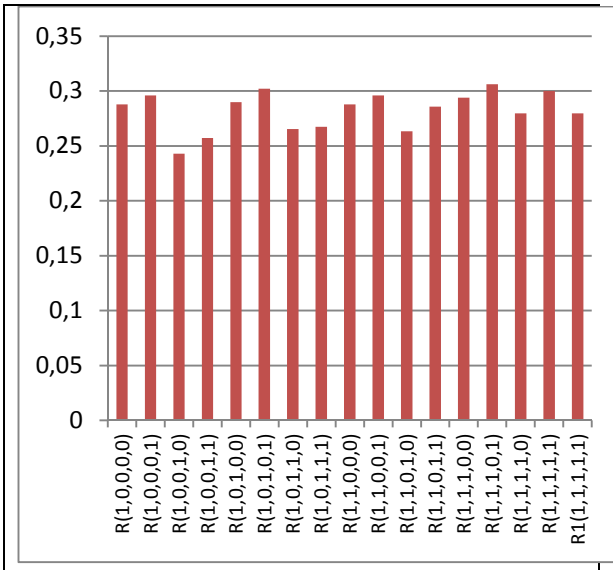


Figure 17: Comparaison graphique de la précision pour les différentes relations sur la collection TREC7

Les 16 combinaisons de R qui sont testées montrent d'une part, que plus on augmente le types des relations pour le calcul de la centralité, plus la précision est améliorée. D'autre part, l'intégration des certains relations diminue la précision, cela est dû à l'étude menée précédemment et qui a montré que les concepts utilisés pour représenter le sens d'un terme ne sont pas les meilleurs.

Plus précisement, sur les deux collections, l'emploi de la relation *est-une-partie-de* avec la relation *synonymie* représenté par R(1,0,1,0,0) a un effet important sur les résultats. Ensuite, l'emploi de la relation *domaine* cumulée avec les deux autres relations comme représenté par R(1,0,1,0,1) amène une amélioration à tous les niveaux de précisions. Enfin, l'emploi de toutes les relations possibles R(1,1,1,1,1) retourne les meilleurs résultats.

D'autre part, l'emploi des relations telles que *dérivé* employés seuls avec la synonymie R(1,0,0,1,0) n'est pas très significatif, tandis que l'emploi de cette relation avec la relation *est-une-partie-de* R(1,0,1,1,0) ramène une amélioration à cette dernière. De sa part, l'emploi de la relation *spécialisation* en combinaison avec d'autres relations diminue la précision globale.

En fait, chacune de ces relations a une spécialité d'améliorer le calcul de la centralité et ainsi augmenter la précision. Mais, toujours on revient à un problème essentiel dans le calcul des relations c'est l'incomplétude de WordNet dans la detection de toutes les relations existantes entre les concepts.

5.8.2 Impact des relations WordNet pour représenter la fréquence conceptuelle

Dans cette section, nous étudions l'impact de chacune des relations WordNet présente dans $R(S, I, P, D, R)$ sur le calcul de la fréquence conceptuelle d'un concept dans un document. Ces expérimentations sont réalisées sur les deux collections « TREC1 » et « TREC7 ». Nous utilisons la même configuration utilisée dans la section précédente, nous utilisons les termes simples et les méthodes de désambiguïsation basées sur la centralité.

Nous étudions l'impact des différentes relations utilisées dans le calcul de *Cf* sur trois fonctions. La première est celle qu'on a proposée en remplaçant la fréquence d'un concept *f*

par sa fréquence conceptuelle *Cf*. La seconde, est la fonction de BM25 en remplaçant *tf* par *Cf*. La troisième, c'est une fonction du modèle de langue qui est utilisé en remplaçant *tf* par *Cf*.

Nous commençons par une description des fonctions que nous allons utiliser dans les expérimentations. Ansi, nous donnons à chaque fonction un nom qui va nous servir à l'identifier par la suite.

La première, on la note « c²*Cf*s » correspond à notre fonction d'appariement en remplaçant *f* par *Cf* dans la fonction pour avoir la fonction suivante :

$$RSV(q,d) = {c_{t,d}}^2 \times Cf_{t,d} \times S_t$$

Avec et $Cf_{t,d}$ (resp. $c_{t,d}$) est la fréquence conceptuelle (resp. la centralité) du mot t dans le document *d*, et S_t est la spécificité de *t* dans WordNet.

Dans la seconde nous remplaçons *tf* dans la formule de BM25 par *Cf*. Elle sera notée « BM25_Cf » et la nouvelle fonction que nous utilisons est la suivante :

$$W(t,d) = \frac{Cf_{t,d} * (k_1 + 1)}{k_1 * ((1-b) + b * \frac{dl}{avdl}) + Cf_{t,d}} \text{x} \log(\frac{N - df(t,C) + 0.5}{df(t,C) + 0.5})$$

Avec et $Cf_{t,d}$ est la fréquence conceptuelle du mot t dans le document *d*, *dl* est la taille du document (le nombre total d'occurrences de mots). $df(t,C)$ est le nombre de document de la collection *C* contenant *t*. k_1 et *b* sont des constantes qui dépendent des collections des tests.

Dans la troisième *tf* est remplacée par *Cf* dans une formule de modèle de langue. Nous nontons cette nouvelle formule « ML_Cf » avec la formule de modèle de langue que nous utilisons, celle qui correspond au Lissage d'indépendance Dirichlet sera la suivante :

$$P(w_i|\, D) = \frac{Cf(w_i, D) + \mu P_{ML}(w_i|C)}{dl + \mu}$$

Avec $P(w_i|\, D)$ représente la probabilité d'un mot w_i dans un document *D*. *dl* est la taille du document (le nombre total d'occurrences de mots), et $Cf(w_i, D)$ est la fréquence conceptuelle du mot w_i dans le document *D*, μ est une constante et $P_{ML}(w_i|C)$ représente la probabilité du w_i dans la collection C qui est estimée par :

$$P_{ML}(w_i|C) = \frac{tf(w_i, C)}{\sum_{\forall\, w_j \in C} tf(w_j, C)}$$

Où et $tf(w_i, C)$ est la fréquence du mot w_i dans la collection *C*, $\sum_{\forall\, w_j \in C} tf(w_j, C)$ représente la somme de fréquences de tous les termes dans la collection *C*.

Pour chacune de ces formules nous varions la combinaison des relations dans R. Les détails des résultats obtenus sont présentés en annexe 3.3. Nous présentons dans le Tableau 13 les résultats simplifiés qui correspondent aux valeurs de la précision pour les top-10 documents pour chaque combinaison des relations possibles de R. Dans la première colonne nous présentons les différentes combinaisons de R. La seconde colonne (resp. la troisième colonne) représente les résultats obtenus par les différentes fonctions (c²*cf*s, BM25_Cf et ML_Cf) appliquées sur la collection TREC1 (resp. la collection TREC7). Dans ce tableau, chaque croisement (ligne, colonne) correspond à une expérimentation menée sur une collection, en appliquant une formule pour une combinaison de certains types de relation. Nous représentons en gras les meilleurs résultats obtenus pour chaque fonction par rapport à la combinaison des relations.

	TREC1			TREC7		
	C²*cf*s	BM25_Cf	ML_Cf	C²*cf*s	BM25_Cf	ML_Cf
R(1,0, 0,0,0)	0,4896	0,4354	0,4771	0,2714	0,3755	0,3755
R(1,0, 0,0,1)	0,4917	0,4333	0,4729	0,2592	0,3755	0,3735
R(1,0, 0,1,0)	0,4896	0,4438	0,4813	0,2592	0,3816	0,3694
R(1,0, 1,0,0)	0,4896	0,4333	0,4750	0,2714	0,3796	0,3796
R(1,1, 0,0,0)	0,4833	0,4729	0,4833	**0,2755**	0,3776	0,3755
R(1,0, 0,1,1)	0,4896	0,4438	0,4792	0,2510	0,3816	0,3673
R(1,0, 1,0,1)	0,4937	0,4313	0,4708	0,2592	0,3796	0,3776
R(1,0, 1,1,0)	0,4875	0,4458	0,4813	0,2612	0,3816	0,3694
R(1,0, 1,1,1)	0,4917	0,4438	0,4771	0,2510	0,3816	0,3673
R(1,1, 0,0,1)	0,4875	0,4542	0,4833	0,2653	0,3796	0,3755
R(1,1, 0,1,0)	0,4896	0,4729	0,4792	0,2694	0,3796	**0,3796**
R(1,1, 0,1,1)	0,4917	0,4646	0,4812	0,2571	0,3816	**0,3796**
R(1,1, 1,0,0)	0,4875	0,4729	0,4833	**0,2755**	0,3755	0,3755
R(1,1, 1,0,1)	0,4854	0,4542	0,4833	0,2633	0,3776	0,3755
R(1,1, 1,1,0)	**0,4938**	0,4708	0,4812	0,2673	0,3755	0,3755
R(1,1, 1,1,1)	0,4896	0,4625	0,4833	0,2551	0,3776	0,3755
R1(1,1, 1,1,1)	0,4813	**0,4896**	**0,5104**	0,2633	**0,3980**	0,3776

Tableau 13: Impact des relations de Cf sur les top-10 documents obtenus sur différentes fonctions et pour les deux collections TREC1 et TREC7

A partir des résultats obtenus sur chacune des deux collections et pour toutes les fonctions nous pouvons constater que chaque relation de *R* influence le calcul de *Cf*. En effet, le fait d'ajouter des relations ça implique que la valeur de Cf augmente et par la suite les résultats s'améliorent. L'homogénéité dans les résultats obtenus dans les différentes expérimentations nous permet de conclure l'intérêt de chaque relation. En particulier, la combinaison de la relation de synonymie avec la relation de hyperonymie (ou est-un) a une grande influence sur le calcul de *Cf*. Ainsi que la combinaison des relations de synonymie, hyperonymie (est-un) et hyponymie (partie-de) est de même intéressante dans le calcul de *Cf*. D'autre part, les relations de domaines, dérivé et hyponymie prisent seul avec la relation de synonymie n'ont pas un grand intérêt dans le calcul du *Cf*.

L'intégration de toutes les relations dans R constitue la meilleure combinaison pour le calcul de *Cf*. Ainsi que la combinaison de toutes les relations indirectement à un premier niveau à partir des concepts qui n'existent pas dans le document représenté par *R1* constitue les meilleurs résultats.

Pour conclure, une meilleure façon de détection des relations entre les concepts d'un document amène à une meilleure valeur de *Cf* ainsi à une forte amélioration dans la précision ramenée par un système de recherche d'information. Plusieurs paramètres influencent le calcul du *Cf*, d'une part une méthode efficace de désambiguïsation des concepts pour détecter le vrai sens d'un concept dans un document. D'autre part une ressource terminologique complète (dans notre expérimentation nous sommes basées sur WordNet) qui prend en compte les différents types de relations avec une grande couverture du lexique de la langue utilisée.

5.9 L'efficacité de la centralité et de la fréquence conceptuelle en RI

Notre objectif dans cette section est d'évaluer l'impact du facteur Cf quand il est intégré dans d'autres modèles de RI performants, à savoir BM25 est un modèle de langue qui correspond au lissage d'indépendance Dirichlet.

Les expérimentations sont réalisées sur les deux collections TREC1 et TREC7. La même configuration d'indexation est utilisée, celle qui correspond aux termes simples et aux méthodes de désambiguïsation des concepts documents et requêtes basés sur la centralité. Ainsi, pour le calcul de la centralité et la fréquence conceptuelle, nous considérons les relations représentées par R1(1,1,1,1,1).

Nous commençons par comparer les résultats obtenus par $c^2*(1+f)*s$ et de Comb(Okapi, c^2*f*s) à ceux d'Okapi-BM25. Ensuite nous comparons les résultats de BM25 basés sur *tf* à ceux basés sur *Cf*. Enfin, nous comparons les résultats obtenus par le lissage de Dirichlet basé sur *tf* à ceux basés sur *Cf*.

Dans le Tableau 14 les fonctions « BM25_tf » correspond à Okapi-BM25 et $c^{2*2}f*s$ correspond $c^2*(1+f)*s$ et BM_c^2*f*s= Comb(Okapi, c^2*f*s), « %imp » représente le pourcentage de gain des résultats obtenus par la fonction F3 à ceux obtenus par la fonction de base F1 pour chaque niveau de précision (P5, P10…). Dans la partie gauche du tableau «Tableau 14 (a) » nous présentons les résultats obtenus sur la collection TREC1, ainsi qu'à droite « Tableau 14 (b) » nous présentons les résultats obtenus sur la collection TREC7.

	BM25_tf	c^2*f*s	BM_c^2*f*s	%imp
P5	0,4500	0,5375	0,5542	23,16%
P10	0,4437	0,4917	0,4958	11,74%
P15	0,4375	0,4861	0,4944	13,01%
P20	0,4323	0,4635	0,4719	9,16%
P30	0,4194	0,4417	0,4486	6,96%
P100	0,3633	0,3633	0,3633	0,00%
Map	0,0930	0,0973	0,0994	6,88%

(a) Comparaison sur TREC1

	BM25_tf	c^2*f*s	BM_c^2*f*s	%imp
P5	0,4204	0,2939	0,4980	18,46%
P10	0,3755	0,2796	0,4408	17,39%
P15	0,3374	0,2667	0,4136	22,58%
P20	0,3112	0,2531	0,3827	22,98%
P30	0,2776	0,2456	0,3211	15,67%
P100	0,1988	0,1988	0,1988	0,00%
Map	0,1247	0,1062	0,1487	19,25%

(b) Comparaison sur TREC7

Tableau 14: Résultats de c^2*f*s

Les résultats obtenus dans le Tableau 14 montrent que sur les deux collections en même temps l'emploi de la combinaison de BM25 avec notre fonction $c^2*(1+f)*s$ augmente la précision de plus de 10% pour le top5, top10 et top15 documents. Ces résultats montrent l'intérêt de cette fonction comparé à BM25, ensuite l'intérêt de l'utilisation de la centralité pour rendre plus des documents pertinents en réponse à une requête.

Dans le Tableau 15 nous comparons les résultats obtenus présentés dans la seconde colonne noté par BM25_tf, à ceux de BM25 basés sur *Cf* représentée par BM25_Cf= « Okapi-BM25 basée sur Cf » dans la troisième colonne. La colonne « %imp » représente le pourcentage de gain des résultats obtenus par BM25_Cf par rapport à BM25_tf pour les top-x niveaux de précision (P5, P10…). A gauche du tableau « Tableau 15 (a) » sont présentés les résultats obtenus sur la collection TREC1, ainsi qu'à droite « Tableau 15 (b) » sont présentés les résultats obtenus sur la collection TREC7.

	BM25_tf	BM25_Cf	%imp
P5	0,4500	0,4875	8,33%
P10	0,4437	0,4896	10,34%
P15	0,4375	0,4778	9,21%
P20	0,4323	0,4604	6,50%
P30	0,4194	0,4625	10,28%
P100	0,3633	0,3633	0,00%
Map	0,0930	0,1043	12,15%

(a) Comparaison sur TREC1

	BM25_tf	BM25_Cf	%imp
P5	0,4204	0,4245	0,98%
P10	0,3755	0,398	5,99%
P15	0,3374	0,3483	3,23%
P20	0,3112	0,3276	5,27%
P30	0,2776	0,2939	5,87%
P100	0,1988	0,1986	-0,10%
Map	0,1247	0,1369	9,78%

(b) Comparaison sur TREC7

Tableau 15: Résultats d'Okapi-BM25 basé sur Cf

Les résultats obtenus dans Tableau 15 présente une amélioration significative pour la précision obtenue sur la collection TREC1 et malgré que les améliorations ne soient pas très intéressantes sur la collection TREC7, nous pouvons constater que l'emploi de Cf dans BM25 améliore les résultats.

Les résultats obtenus en remplaçant *tf* par *Cf* dans la fonction du modèle de langue sont présentés dans le Tableau 16. Les deux tableaux représentent les résultats obtenus sur la collection TREC1 dans la partie (a) (resp. sur la collection TREC7 dans la partie (b)). Dans la seconde colonne de chacun des 2 tableaux nous représentons les résultats obtenus par le lissage de Dirichlet du modèle de langue noté par ML_tf. Dans la troisième colonne nous remplaçons *tf* par *Cf* dans cette dernière fonction que nous notons par ML_Cf= « Lissage de Dirichlet Basé sur Cf». Les pourcentages d'amélioration obtenus par l'emploi de Cf sont représentés dans la quatrième colonne.

	ML_tf	ML_Cf	%imp
P5	0,5000	0,5542	10,84%
P10	0,4708	0,5104	8,41%
P15	0,4750	0,4958	4,38%
P20	0,4448	0,4917	10,54%
P30	0,4264	0,4625	8,47%
P100	0,3633	0,3633	0,00%
Map	0,0966	0,1060	9,73%

(a) Comparaison sur TREC1

	ML_tf	ML_Cf	%imp
P5	0,3633	0,4204	15,72%
P10	0,3184	0,3776	18,59%
P15	0,2952	0,3483	17,99%
P20	0,2694	0,3235	20,08%
P30	0,2476	0,2898	17,04%
P100	0,1988	0,1986	-0,10%
Map	0,1144	0,1354	18,36%

(b) Comparaison sur TREC7

Tableau 16: Résultats du lissage de Dirichlet basé sur Cf

Les résultats obtenus dans le Tableau 16 justifient les résultats obtenus dans l'expérimentation précédente que l'emploi de Cf en remplaçant tf améliore la performance des recherches. Ces améliorations correspondent à un pourcentage supérieur à 5% à tous les niveaux de précision.

Les résultats obtenus par les différentes fonctions et sur différentes collections montrent que la précision obtenue en employant la centralité et la fréquence conceptuelle est meilleure à tous les niveaux que l'emploi de *tf*. La troisième colonne de différents tableaux (Tableau 14, Tableau 15 et Tableau 16) montre une amélioration particulière supérieur de 5% dans les tops documents ce qui signifie que les résultats sont intéressants et ils ne sont pas obtenus par hasard.

Les deux paramètres $Cf_{t,d}$ et $c_{t,d}$ d'un terme t dans un document d sont deux nouveaux facteurs intéressants pour la recherche d'information qui doivent former un passage entre

l'aspect fréquentiel vers un aspect sémantiquement fréquentiel pour pondérer un document par rapport à une requête.

5.10 Conclusion

Dans ce chapitre nous avons introduit deux nouveaux facteurs permettant de mesurer l'intérêt ou l'importance d'un terme, la centralité et la fréquence conceptuelle. Ces deux facteurs permettent de rendre compte du nombre de liens sémantiques que peut avoir un terme avec les autres termes d'un document. Un terme est d'autant plus représentatif du contenu d'un document qu'il est sémantiquement relié aux autres termes du document.

Ces facteurs ont été utilisés dans deux tâches. Tout d'abord pour désambiguïser les concepts d'un document, ensuite pour mesurer la pertinence d'un document vis-à-vis d'une requête. Nous avons effectué une série d'expérimentations dans lesquelles nous avons montré la prise en compte de ces facteurs pour améliorer significativement les performances en RI.

Conclusion Générale

Synthèse

Dans ce mémoire, nous nous situons dans le contexte de l'accès sémantique aux informations. Plus précisément, nous proposons de nouveaux paramètres pour identifier les termes significatifs d'un texte. Ces paramètres vont au-delà du simple comptage des termes dans leur contexte d'apparition désigné par *tf* dans la recherche d'information classique. Ils consistent à mesurer un degré de centralité d'un terme *t* dans un document *d* qui est estimé en fonction des termes de *d* qui sont en relation directe avec *t*.

La centralité constitue le sujet d'étude dans ce mémoire. Dans notre contribution, nous avons cherché la meilleure façon de calculer sa valeur, ainsi nous avons définis d'autres facteurs sémantiques qui influent cette valeur. Par la suite, la centralité a été employée dans différentes propositions soit pour extraire le contenu sémantique d'un texte, soit pour identifier les documents pertinents en réponse à une requête.

Pour calculer la centralité d'un terme dans un document *d*, l'étape essentielle consiste à extraire pour le document *d* ces différents concepts, nous déterminons pour chaque terme identifié dans le document un concept qui lui correspond dans une ontologie *O* des concepts. Ensuite, identifier les liens entre ces concepts en exploitant les relations sémantiques existantes dans l'ontologie *O* pour construire ce qu'on appelle 'cluster conceptuel. A l'intérieur de chaque cluster, chaque concept *t* a un niveau de "centralité" calculé de deux manières. D'une part, nous estimons la *centralité* de *t* dans *d* par le nombre de concepts dans *d* qui sont en relations sémantiques avec *t*. D'autre part, nous estimons la *fréquence conceptuelle* d'un concept *t* dans le document *d* par la somme des fréquences des concepts qui sont en relations sémantiques directes avec *t*.

L'intérêt de la centralité peut être perçu dans le fait qu'on peut estimer un poids pour un terme par rapport à un document même si ce dernier n'apparaît pas dans le document. Comparé à *tf* (term frequency) qui affecte un poids à un terme *t* par rapport à un document *d* en fonction de son nombre d'apparition dans *d*, la centralité affecte un poids à *t* même s'il n'apparaît pas dans *d* en fonction des termes de *d* qui lui sont sémantiquement reliés. Cette proposition est intéressante dans la recherche d'information pour pondérer les termes d'une requête par rapport à un document indépendamment de la présence des termes de la requête dans le document recherché.

Pour évaluer l'intérêt de la centralité, nous l'avons employé pour répondre aux différentes problématiques évoquées dans l'introduction générale. La première concerne le problème de la polysémie des unités langagières que représentent les mots dans un texte, nous avons proposé une méthode désambiguïsation détectant le meilleur sens des éléments (mots clés, groupes de mots) présents dans un texte. Nous nous sommes basés sur une ontologie générale (dans notre cas, nous nous sommes basés sur WordNet) pour identifier les concepts suceptibles de représenter les différents sens de chaque élément identifié. Dans notre approche, nous avons employés la centralité pour représenter chaque élément du texte par un unique concept. Dans le chapitre 5, la méthode proposée a été comparée à d'autres méthodes de désambigïsation qui ont montré son efficacité.

La seconde approche présentée dans le chapitre 4 correspond à une méthode pour l'extraction de point de vue sémantique du contenu du texte, indépendamment de toute requête, qui va au-delà de l'indexation statistique standard. Dans cette méthode, nous avons représenté les textes par une vue multi-niveau des mots permettant de donner un aperçu hiérarchique des sujets présentés dans le document. Les mots significatifs sont représentés par des groupes sémantiques flous pondérés des mots sur plusieurs niveaux. Chaque niveau inférieur contient des mots de moins en moins significatifs et donne plus de détails sur le document. Tout d'abord, les mots importants et significatifs sont identifiés à partir des clusters et leurs critères (la fréquence, la centralité, la spécificité…), où différentes fonctions d'agrégations floues de ces critères ont été testées. Ensuite, quelques phrases contenant un maximum de ces mots sont extraites du texte et sont proposées comme représentatives de son contenu. Cette méthode ainsi que son intérêt sont testés sur des textes de longueurs différentes. Dans les différentes expérimentations menées au chapitre 4, nous avons d'abord montré l'efficacité de notre méthode pour extraire les titres d'un texte où environ 74% des titres des documents de tests ont été identifiés. Ensuite, l'intérêt de ces mots représentatifs est testé dans l'identification des phrases représentatives d'un document contenant les mots significatifs identifiés pour former un résumé textuel où environ 83% des thèmes de documents de la collection sont identifiés.

Dans la troisième approche comme proposée dans le chapitre 5, nous avons introduit les nouveaux facteurs proposés *la centralité* et *la fréquence conceptuelle* pour mesurer la pertinence d'une requête vis-a-vis d'un document dans le domaine de la recherche d'information. Dans ce cadre, nous avons employé ces facteurs dans différentes fonctions de pondération du document vis-à-vis d'une requête. Dans un premier temps, nous avons proposé une fonction d'appariement qui est une combinaison de ces facteurs avec d'autres facteurs tels que la *spécificité* et la *fréquence*. Ensuite, nous avons intégré ces facteurs dans des fonctions d'appariement connues en remplaçant *tf* (la fréquence usuelle d'un terme) par la *fréquence conceptuelle*. Enfin, vu que les documents pertinents retirés en premier par la méthode BM25 sont différents de ceux retirés par notre fonction, nous avons proposés une fonction combinant notre fonction avec la fonction de BM25. Pour évaluer l'intérêt des différents facteurs en RI, nous avons comparé les résultats obtenus par notre méhtode de pondération à ceux obtenus par des modèles connus tels que la fonction de modèle de langue qui correspond au lissage d'indépendance de Dirichlet et la fonction OKAPI-BM25. Les différents résultats obtenus ont montré que l'intégration des facteurs sémantiques que nous avons proposés ramène à une amélioration au niveau de précision dans un moteur de recherche d'information. En plus de ces expérimentations, nous avons mené des expériences sur les meilleures relations sémantiques permettant de calculer la centralité et la fréquence conceptuelle.

Perspectives

Dans notre proposition, nous avons remplacé la vue statistique pure d'un document par une vue tirant avantage des critères sémantiques tels que la centralité, la fréquence conceptuelle et la spécificité. L'efficacité de cette représentation a été employée dans des expérimentations de tests pour extraire à partir d'un texte, les termes intéressants ainsi que pour représenter son contenu par un ensemble des phrases significatives. Ces expérimentations doivent être élargies sur une plus large échelle en introduisant différentes fonctions d'agrégation de ces critères.

La méhtode d'extraction des phrases représentatives d'un document doit être évaluée. D'une part, elle doit être intégrée à un moteur de recherche pour représenter le contenu d'un snippet. Cette prespective consiste à proposer un protocole d'évaluation qui fait intervenir des évaluateurs humains pour juger la performance de ces résumés. D'autre part, cette méhtode doit être évaluée dans l'extraction des résumés textuels d'un document.

Les mêmes critères ont été introduits en recherche d'information en proposant une fonction d'appariement simple. Une prespective de cette proposition sera de mener une étude complète sur la distribution de différents paramètres dans les documents pertinents et non pertinents à une requête pour obtenir une meilleure fonction d'appariement pour la recherche d'information.

Les critères que nous avons proposés sont basés sur la centralité d'un terme dans un document c.à.d. le nombre des termes d'un document qui sont en relation avec ce terme. Pour montrer l'intérêt de notre proposition, nous nous sommes basés sur WordNet. L'intérêt de cette notion de centralité peut être amélioré de différentes manières, d'une part on peut considérer uniquement la centralité d'un terme par rapport à une partie du document (peut être un paragraphe, ensemble de paragraphe, une section…). D'autre part cette centralité peut être améliorée en intégrant différentes ressources terminologiques pour la détection des relations entre les termes.

D'autres perspectives peuvent être ouvertes en intégrant des ontologies de domaines. L'intégration des ontologies de domaines peut avoir plusieurs intérêts. D'une part, on peut préciser pour un document le domaine qui lui correspond, ainsi que ces ontologies peuvent être employées pour calculer la centralité d'un terme dans un document par rapport à un domaine. Dans ce cas, ces facteurs seront plus efficaces pour représenter une recherche propre à un domaine.

REFERENCES BIBLIOGRAPHIQUES

Bibliographie

[Aggarwal et al, 2001] Charu C. Aggarwal and Philip S. Yu. On Effective Conceptual Indexing and Similarity Search in Text Data. In Proceedings of ICDM '01, the 2001 IEEE International Conference on Data Mining. Washington, DC, USA, 2001.

[Agirre et al, 2000] E. Agirre and M. David. ``Exploring automatic word sense disambiguation with decision lists and the Web" In: Proceedings of the Semantic Annotation And Intelligent Annotation workshop organized by COLING, Luxembourg, 2000.

[Agirre et al, 2001] E. Agirre, A. Olatz, M. David, and H. Eduard. ``Enriching WordNet concepts with topic signatures." In: Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources, Pittsburgh, June 2001. http://www.seas.smu.edu/~rada/mwnw/papers/WNW-NAACL-228.pdf.gz

[Agrawal et al, 1998] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications. In Proceedingsof the International Conference on Management of Data, (SIGMOD), volume 27(2) of SIGMOD Record, pages 94–105, Seattle,WA, USA, 1–4 June 1998. ACM Press.

[Ahuja et al, 1993] R.K. Ahuja, T.L. Magnanti, J.B. Orlin, Network Flows : Theory, Algorithms, and Applications. Prentice Hall, Englewood Cliffs, New Jersey, 1993 (ISBN 0-13-617549-X).

[Anderberg, 1973] Michael R. Anderberg. Cluster analysis for applications. Academic Press, 1973.

[Andreasen et al. 2003] T. Andreasen: An approach to Knowledge-based Query Evaluation. in special issue on "Fuzzy Databases" in Fuzzy Sets and Systems, ISSN 0165-0114 Publisher: Elsevier Science 2003.

[Andreasen et al, 2002] T. Andreasen, P. A. Jensen, J. F. Nilsson, P. Paggio, B. S. Pedersen, H. E. Thomsen: Ontological Extraction of Content for Text Querying. NLDB 2002: 123-136.

[Anil et al, 1988] K. J. Anil and C. D. Richard. Algorithms for Clustering Data. Prentice-Hall, 1988.

[Ankerst et al, 1996] M. Ankerst, M. M. Breunig, H.P. Kriegel, and J. Sander. OPTICS: Ordering Points To Identify the Clustering Structure. In Proceedings of the International Conference on Management of Data, (SIGMOD), volume 28(2) of SIGMOD Record, pages 49–60, Philadelphia, PA, USA, 1–3 June 1996. ACM Press.

[Anthonisse, 1971] J. Anthonisse, “The Rush in a Graph,” Amsterdam: University of Amsterdam Mathematical Centre, 1971.

[Aone et al, 1999] C. Aone, M. E. Okurowski, J. Gorlinsky, et B. Larsen. A Trainable Summarizer with Knowledge Acquired from Robust NLP Techniques, 71–80. MIT Press. 1999.

[Aussenac-Gilles et al, 2000] N. Aussenac-Gilles, B. Biébow, N. Szulman, Revisiting Ontology Design: a method based on corpus analysis. Knowledge engineering and knowledge management: methods, models and tools, Proc. of the 12th International Conference on Knowledge Engineering and Knowledge Management. Juan-Les-Pins (F). Oct 2000. R Dieng and O. Corby (Eds). Lecture Notes in Artificial Intelligence Vol 1937. Berlin: Springer Verlag. pp. 172-188. 2000.

[Baeza-Yates & Ribeiro-Neto, 1999] R. Baeza-Yates and B. Ribeiro-Neto (1999), Modern Information Retrieval. Addison-Wesley. ISBN 0-201-39829-X.

[Ballmer, 1976] T. T. Ballmer, Fuzzy punctuation or the continuum of grammaticality. Memo ERLM590, Univ. of California, Berkeley, 1976.

[Banerjee et al, 2003] S. Banerjee and T. Pedersen. ``Extended Gloss Overlaps as a Measure of Semantic Relatedness" In: Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence IJCAI-2003, Acapulco, Mexico, August, 2003. http://www.d.umn.edu/~tpederse/Pubs/ijcai03.pdf

[Barry, 1994] C. L. Barry. User-defined relevance criteria : an exploratory study. Journal of the American Society for Information Science, 45 :149–159, 1994.

[Basili et al, 1997] R. Basili, M. DellaRocca and M. T. Pazienza. ``Contextual word sense tuning and disambiguation." In: Applied Artificial Intelligence 11 (3), 1997, pp. 235 – 262.

[Bausch et al, 2006] P. Bausch, J. Bumgardner. "Make a Flickr-Style Tag Cloud", Flickr Hacks. O'Reilly Press. ISBN 0596102453. 2006.

[Baziz et al, 2004] M. Baziz, M. Boughanem, G. Pasi, H. Prade. A fuzzy set approach to concept-based information retrieval. Proc.10th.IPMU, Perugia, 2004, 1775-1782.

[Baziz, 2005] M. Baziz Indexation conceptuelle guidée par ontologie pour la recherche d'information. Thèse de doctorat, Université Paul Sabatier, décembre 2005.

[Baziz et al, 2006] M. Baziz, M. Boughanem, G. Pasi, H. Prade, A fuzzy logic approach to information retrieval using an ontology-based representation of documents. In: Fuzzy Logic and the Semantic Web (E. Sanchez, ed.), Elsevier, 363-377, 2006.

[Beigberder et al, 2004] M. Beigberder, Annabelle Mercier. Application de la logique floue à un modèle de recherche d'information basé sur la proximité. in Actes des 12es rencontres francophones sur la Logique Floue et ses Applications Nantes, France, pp. 231--237, 2004.

[Belkin et al, 1992] N. J. Belkin, P. Ingwersen, A. M. Pejtersen: Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Copenhagen, Denmark, June 21-24, 1992 ACM 1992.

[Berge, 1985] C. Berge. Graphes. Gauthier-Villars, 1985 (ISBN 2-04-015555-4).

[Berge, 1987] C. Berge, Hypergraphes. Gauthier-Villars, 1987 (ISBN 2-04-016906-7).

[Berkhin, 2002] P. Berkhin (2002). Survey of clustering data mining techniques. Technical report, Accrue Software, San Jose, California.

[Bielenberg et al, 2006] K. Bielenberg, and M. Zacher. Groups in Social Software: Utilizing Tagging to Integrate Individual Contexts for Social Navigation, Masters Thesis submitted to the Program of Digital Media, Unisersitat Bremen. 2006.

[Bizer et al, 2009] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. Dbpedia - a crystallization point for the web of data. J. Web Sem., 7(3) :154–165, 2009.

[Black, 2006] P. E. Black. "Manhattan distance", in Dictionary of Algorithms and Data Structures[online] chap2,ed., U.S. National Institute of Standards and Technology. 31 May 2006. Paul E (http://www.itl.nist.gov/div897/sqg/dads/).

[Bonacich, 1972] P. Bonacich. “Factoring and weighting approaches to status scores and clique identification,” Journal of Mathematical Sociology, vol. 2, no. 1, pp. 113–120, 1972.

[Boubekeur et al, 2008] F. Boubekeur, M. Boughanem, L. Tamine. Une approche d’indexation conceptuelle de documents basée sur les graphes CP-Nets. Dans : Colloque sur l'Optimisation et les Systèmes d'Information (COSI 2008), Tizi-Ouzou (Algérie), 08/06/2008-10/06/2008, Université Mouloud Mammeri, p. 439-450, juin 2008.

[Boughanem et al, 2003] M. Boughanem, K. Sauvagnat, C. Laffaire. (2003). Mercure at TREC'2003, Web track - Topic Distillation Task. Text REtrieval Conference (TREC 2003). Gaithersburg, Maryland, USA, 18-21 /11/2003, National Institute of Standards and Technology (NIST).

[Boughanem et al, 2010] M. Boughanem, I. Mallak, H. Prade. A new factor for computing the relevance of a document to a query (regular paper). Dans : IEEE World Congress on Computational Intelligence (WCCI 2010), BARCELONE, 18/07/2010-23/07/2010, 2010.

[Bouidghaghen et al, 2008] O. Bouidghaghen, M. Boughanem, H. Prade. Extraire les thématiques des textes: Vers une approche par la logique floue. Actes Rencontres Francophones sur la Logique Floue et ses Applications (LFA 2008), Lens, 16-17/10/2008, Cépaduès Editions, 34-41, 2008.

[Bouidghaghen et al, 2009] O. Bouidghaghen, M. Boughanem, H. Prade, I. Mallak. (2009). A fuzzy logic approach to topic extraction in texts. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 17, N. Suppl. 1, 81-112.

[Bourigault, 1994] D. Bourigault. Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases, dans Proceedings of the Fourteenth International Conference on Computational Linguistics-COLING 92, Nantes, p. 977-981.

[Bourigault, 1996] D. Bourigault. (1996). " Lexter, a Natural Language Processing Tool for Terminology Extraction ". Proceedings of Euralex’96, Göteborg University, Department of Swedish, 1996, pp. 771-779.

[Brachman et al,1985] R.J. Brachman, J.G. Schmolze. An Overview oft he KL_ONE Representation System. In Cognitive Science, 1985, Vol. 9. P 171-216.

[Brachman, 1983] R.J. Brachman, R.E. Fikes, H.J. Levesque. Krypton: A Functional Approach to Knowledge Representation. In Computer, 1983, Vol. 16, N°10 p 67-73.

[Brandes, 2001] U. Brandes. A faster algorithm for betweenness centrality. Journal of Mathematical Sociology http://www.cs.ucc.ie/~rb4/resources/Brandes.pdf, 2001.

[Brandow et al, 1995] R. Brandow, K. Mitze, et L. F. Rau. Automatic condensation of electronic publications by sentence selection. Information Processing and Management : an International Journal 31(5), 675–685. 1995.

[Brin et Page, 1998] S. Brin et L. Page. The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems 30(1-7), 107–117. 1998.

[Buckley et al, 2000] C. Buckley and E. M. Voorhees. Evaluating evaluation measure stability. In SIGIR '00 : Proceedings of the 23rd annual international ACM SIGIR on Research and development in information retrieval, pages 33–40, New York, NY, USA, 2000. ACM.

[Buell et al, 1981] D. A. Buell, and D. H. Kraft. A model for a weighted retrieval system. J. of American Society for Informat. Sci., 32, 211-216, 1981.

[Candillier, 2006] L. Candillier. Thèse « Contextualisation, visualisation et évaluation en apprentissage non supervisé » de Laurent Candillier (Université de Lille 3), 2006/09/15, PDF, 250 pages.

[Shannon, 1948] Claude Shannon. A mathematical theory of communication Bell System Technical Journal, vol. 27, pp. 379-423 and 623-656, July and October, 1948.

[Chaumartin, 2007a] F. Chaumartin. (2007). Extraction de paraphrases désambiguïsées à partir d'un corpus d'articles encyclopédiques alignés automatiquement. Actes de TALN 2007.

[Chaumartin, 2007b] F.R. Chaumartin. WordNet et son écosystème : un ensemble de ressources linguistiques de large couverture. BDL-CA, Montréal, 23 avril 2007.

[Chevallet, 1992] J-P Chevallet. Un modèle logique de recherche d'informations appliqué au formalisme des graphes conceptuels – L prototype ELEN et son expérimentation sur un corpus de composants loficiels. Thèse de doctorat en informatique : Université Joseph Fourier, Grenoble, France, 1992. 202 pages.

[Church & Hanks, 1989] K. W. Church, and P. Hanks. (1989). Word Association Norms, Mutual Information, and Lexicography. In Proceeding of the 27th Annual Meeting of the Association for Computational Linguistics, ACL'89, Vancouver, Canada.

[Church & Hanks, 1990] K. W. Church and P. Hanks. (1990). Word Association Norms, Mutual Information and Lexicography. Computational Linguistics, 16(1), 22-29.

[Church et al, 1991] K. W. Church, W. Gale, P. Hanks and D. Hindle. (1991). Using Statistics in Lexical Analysis. In U. Zernick, Ed., Lexical Acquisition: Using On-line Resources to Build a Lexicon, p. 115-164. Laurence Erlbaum.

[Chotteau, 2003] C. Chotteau. Corrélation smantique entre documents : application à la recherche d'information juriqique sur le Web. PhD thesis, Centre de recherche en informatique, Mines de Paris [ENSMP], Informatique temps réel, robotique et automatique, 2003.

[Clarke et al, 2000] L. A. Clarke Charles, V. Cormack Gordon, and Elizabeth A. Tudhope. Relevance ranking for one to three term queries. Information Processing and Management, 36(2) :291.311, 2000.

[Claveau 2003] V. Claveau. Acquisition automatique de lexiques sémantiques pour la recherche d'information, thèse de doctorat, Université de Rennes 1.

[Cleverdon, 1962] C. W. Cleverdon. Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems. Cranfield, U.K.: College of Aeronautics. Aslib Cranfield Research Project, 1962.

[Cleverdon, 1976] C. W. Cleverdon. The Cranfield tests on index language devices. Aslib Proceedings 19(6), 173-193, 1967.

[Conroy et O'leary, 2001] J. M. Conroy et D. P. O'leary. Text summarization via hidden Markov models. Dans les actes de 24th annual international ACM SIGIR conference on Research and development in information retrieval, 406–407. ACM New York, NY, USA. 2001.

[Croft et al, 1992] J. P. Callan, W. B. Croft, M. Stephen. Harding: The INQUERY Retrieval System. DEXA 1992: 78-83.

[da Cunha et al, 2007] I. da Cunha, S. Fernandez, P. Velázquez, J. Vivaldi, E. SanJuan, et J.-M. Torres-Moreno. A new hybrid summarizer based on Vector Space model, Statistical Physics and Linguistics. Dans les actes de Lecture Notes In Computer Science, Volume 4827, 872. Springer. 2007.

[Daille, 1994] B. Daille. Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques. Rapport interne, Université de Paris 7. Thèse de Doctorat en Informatique Fondamentale.

[Dangalchev, 2006] Ch. Dangalchev. Residual Closeness in Networks, Phisica A 365, 556 (2006).

[Daoud, 2009] M. Daoud. Accès personnalisé à l''information : approche basée sur l''utilisation d'un profil utilisateur sémantique dérivé d''une ontologie de domaines à travers l''historique des sessions de recherche.Thèse de doctorat, Université Paul Sabatier, décembre 2009.

[David et Plante, 1991] S. David et P. Plante. (1991). « Termine v. 1.0™ : principes et propriétés linguistiques » in Actes du colloque Industries de la langue, nov. 1990. Montréal : OLP et Société des traducteurs du Québec. P. 71-88.

[De Loupy et al,2002] C. De Loupy and M. El-Bèze. Managing Synonymy and Polysemy in a Document Retrieval System Using WordNet. In Proceedings of the LREC'02 Workshop on Using Semantics for Information Retrieval and Fitering, Las Palmas de Gan Canaria, Espagne.

[Deerwester et al, 1990] S.C. Deerwester, S. Dumais, T.K. Landauer, G.W. Furnas, R.A. Harshman. Indexing by Latent Semantic Analysis, Journal of the American Society of Information Science, 41(6), p. 391-407, 1990.

[Demant, 1971] B. Demant, Fuzzy-Retrieval-Strukturen, Angew. Inf., Appl. Inf., 13, 500-502, 1971.

[Deza et al, 1994] E. Deza and M. M. Deza. (2009) Encyclopedia of Distances, page 94, Springer.

[Dorr et al, 1996] B.J. Dorr and A. Jones Douglas. ``Acquisition of semantic lexicons: using word sense disambiguation to improve precision.'' In: Proceedings of the ACL SIGLEX Workshop on Breadth and Depth of Semantic Lexicons, Santa Cruz, June 1996. ftp://ftp.umiacs.umd.edu/pub/jones/siglex-96.ps.gz

[Dunning, 1994] T. Dunning. (1994). Statistical identification of language. Computing Research Laboratory Technical Memo MCCS 94-273, New Mexico State University, Las Cruces, New Mexico.

[Edmundson, 1969] H. P. Edmundson. New Methods in Automatic Extracting. Journal of the ACM (JACM) 16(2), 264–285. 1969.

[Erkan et al, 2004] G. Erkan, R. Radev. Dragomir. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. Department of EECS, School of Information University of Michigan, Ann Arbor, MI 48109 USA, 2004.

[Ester et al, 1996] M. Ester, H.P. Kriegel, J. Sander, and X. Xu. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, Oregon, USA: AAAI Press, pp. 226–231.

[Esuli et al, 2006] A. Esuli, F. Sebastiani. (2006). SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining. Actes de LREC 2006, fifth international conference on Language Resources and Evaluation, pp. 417-422.

[Euler, 1741] L. Euler - Solutio problematis ad geometriam situs pertinentis [archive] chap2 [archive] chap2, Commentarii academiae scientiarum Petropolitanae 8, 1741, pages 128-140.

[Favre et al, 2006] B. Favre, F.Béchet, P. Bellot, F. Boudin, M. El-Bèze, L. Gillard, G. Lapalme, J.-M. Torres-Moreno. The LIA-Thales summarization system at DUC-2006, Document Understanding Conference (DUC-2006), New York (USA), 2006.

[Fellbaum et al, 2001] Christiane Fellbaum, Martha Palmer, Hoa Trang Dang, Lauren Delfs, and Susanne Wolf. ``Manual and automatic semantic annotation with WordNet.'' In: Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources, Pittsburgh, June 2001. http://www.seas.smu.edu/~rada/mwnw/papers/invitedPaper.pdf

[Fellbaum, 1998] Ch. Fellbaum. WordNet:An Electronic Lexical Database. Cambridge Massachusettes, Etats-Unis: The MIT Press.

[Fillmore et al, 2002] Ch. J. Fillmore, C. F. Baker, and H. Sato. (2002) “The FrameNet Database and Software Tools”, in Proceedings of the Third International Conference on Language Resources and Evaluation (LREC), Las Palmas: 1157-1160.

[Fillmore, 1968] G. Fillmore. (1968). The case for case. Universals in Linguistic Theory. Bach and Harms. p.1 – 90.

[Fisher, 1987] D.H. Fisher. Knowledge Acquisition via Incremental Conceptual Clustering. Machine Learning 2, 139--172, (1987).

[Freeman, 1977] L. Freeman. “A Set of Measures of Centrality Based on Betweenness,” Sociometry, vol. 40, no. 1, pp. 35–41, 1977.

[Freeman, 1979] L. Freeman. “Centrality in social networks: Conceptual clarification,” Social Networks, vol. 1, no. 3, pp. 215–239, 1979.

[Garey et al, 1979] M.R. Garey, D.S. Johnson. Computer and Intractability : A Guide to the Theory of NP-completeness. W.H. Freeman, San Fransisco, 1979 (ISBN 0-7167-1044-7, 0-7167-1045-5).

[Gentilhomme, 1968] Y. Gentilhomme. Les ensembles flous en linguistique. Cahiers de Linguistique Théorique et Appliquée (Bucharest) 5, 47-63, 1968.

[Girvan et al, 2002] M. Girvan and M. Newman. “Community structure in social and biological networks,” Proceedings of the National Academy of Sciences, vol. 99, no. 12, p. 7821, 2002.

[Gloud, 1991] R.J. Gould. "Updating the Hamiltonian Problem - A Survey", Journal of Graph Theory 15 (1991), 121-157.

[Gonzalo et al, 1998] J. Gonzalo, F. Verdejo, I. Chugur, and J. Cigarrán. Indexing with WordNet synsets can improve text retrieval, in Proc. the COLING/ACL '98 Workshop on Usage of WordNet for Natural Language Processing, 1998.

[Graham et al, 1995] R. Graham, M. Grötschel, and L. Lovász. Handbook of Combinatorics, North-Holland, Amsterdam, 1995, 2 volumes (ISBN 0-444-88002-X).

[Guha et al, 1998] S. Guha, R. Rastogi, and K. Shim. CURE: An Efficient Clustering Algorithm for Large Databases. In Proceedings of the International Conference on Management of Data, SIGMOD),volume 27(2) of SIGMOD Record, pages 73–84, Seattle,WA, USA, 1–4 June 1998. ACM Press.

[Haddad, 2002] M.H. Haddad. Extraction et Impact des connaissances sur les performances des Systèmes de Recherche d'Information, Thèse de Doctorat en Informatique de L'Université Joseph Fourier, Grenoble, 2002.

[Halvey et al, 2007] M. Halvey and M. T. Keane. An Assessment of Tag Presentation Techniques, poster presentation at WWW 2007, 2007.

[Harbeck et al, 1999] S. Harbeck. Uwe Ohler, Elmar Nöth, Heinrich Niemann: Information Theoretic Based Segments for Language Identification. TSD 1999: 187-192.

[Harman, 1993] D. Harman. Overview of TREC-1, Proceedings of the workshop on Human Language Technology, March 21-24, 1993, p. 61-65, Princeton, New Jersey.

[Harrathi, 2009] F. Harrathi. Extraction de concepts et de relations entre concepts à partir des documents multilingues : approche statistique et ontologique. Septembre 2009.

[Hassan-Montero et al, 2006] Y. Hassan-Montero, V. Herrero-Solana. Improving Tag-Clouds as Visual Information Retrieval Interfaces. International Conference on Multidisciplinary Information Sciences and Technologies, InSciT2006. Mérida, Spain. October 25-28, 2006.

[Hawking et Thistlewaite, 1995] D. Hawking, and P. Thistlewaite. Proximity operators - so near and yet so far. In D. K. Harman, editor, TREC-4 proceedings. NIST, 1995.

[Hendrix, 1978] G. Hendrix. The Representation of Semantic Knowledge. In Understanding Spoken Langage. Edited by D.E. Walker, New York: North Holland, 1978. P.121-226.

[Hiemstra et al, 2001] D. Hiemstra, and S. Robertson. "Relevance feedback for best match term weighting algorithms in information retrieval", In Alan Smeaton and Jamie Callan (Eds.) Proceedings of the Joint DELOS-NSF Workshop on Personalisation and Recommender Systems in Digital Libraries, ERCIM Workshop Proceedings 01/W03, pages 37-42, Dublin City University, June 2001

[Hinneburg et al, 1998] A. Hinneburg, and D. A. Keim. An Efficient Approach to Clustering in Large Multimedia Databases with Noise. In Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining, (KDD), pages 58–65, New York, NY, USA, 27–31 August 1998. AAAI Press.

[Hirst et al, 1998] G. Hirst, and D. St Onge. (1998) Lexical chains as representations of context for the detection and correction of malapropisms. In Christiane Fellbaum (editor), WordNet: An electronic lexical database, Cambridge, MA: The MIT Press.

[Huang et al, 2003] Y.-P. Huang, L.-J. Kao. Tienwei Tsai, and Dankai Liu.Using Fuzzy Centrality and Intensity Concepts to Construct an Information Retrieval Model. This paper appears in: Systems, Man and Cybernetics, 2003. IEEE International Conference on Publication Date: 5-8 Oct. 2003

[Huang, 1998] Z. Huang. (1998). Extensions to the K-means Algorithm for Clustering Large Datasets with Categorical Values. Data Mining and Knowledge Discovery, 2, p. 283-304.

[Ingwersen, 1992] P. Ingwersen. Information retrieval interaction. London, Taylor Graham, 1992.

[Jacquemin et al, 2002] C. Jacquemin, B. Daille, J. Royante and X. Polanco. In vitro evaluation of a program for machine-aided indexing. Inf. Process. Manage. 38, 6 (Nov. 2002), 765-792.

[Jacquemin, 1997] C. Jacquemin. Variation terminologique: Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus. Mémoire d'habilitation à diriger des recherches en informatique fondamentale, Université de Nantes.

[Järvelin et al, 2004] E. Airio, K. Järvelin, P. Saatsi, J. Kekäläinen, and S. Suomela. (2004). CIRI - An Ontology-based Query Interface for Text Retrieval. In Hyvönen, E, Kauppinen,

T, Salminen, M, Viljanen, K & Ala-Siuru, P, eds. Web Intelligence. Helsinki: Finnish Artificial Intelligence Society. 73-82.

[Jensen et al, 1995] T.R. Jensen, B. Toft, Graph Coloring Problems. Wiley, New York, 1995 (ISBN 0-471-02865-7).

[Jiang et al, 1997] J. Jiang and D. Conrath. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In Proceedings on International Conference on Research in Computational Linguistics, Taiwan, 1997.

[Joyce et al, 2005] J. Rong, Y. Joyce, and L. Si Chai. (2005). Learn to weight terms in information retrieval using category information. ACM International Conference Proceeding Series; Vol. 119, Proceedings of the 22nd International Conference on Machine Learning (pp. 353 - 360). Bonn, Germany : ACM, New York, NY, USA.

[Justeson & Katz, 1995] J. Justeson, and S. Kartz. Technical Terminology : some linguistic properties and algorithm for identification in text. Natural Language Engineering, 1(1), 9-27.

[Karov et al, 1996] Y. Karov, and E. Shimon. ``Learning similarity-based word sense disambiguation from sparse data." In: Proceedings of the 4th Workshop on Very Large Corpora, Copenhagen, 1996. http://xxx.lanl.gov/abs/cmp-lg/9605009

[Kaser et al, 2007] O. Kaser and D. Lemire. Tag-Cloud Drawing: Algorithms for Cloud Visualization, Tagging and Metadata for Social Information Organization (WWW 2007), 2007.

[Katz et al, 1998] Ö. Uzuner, B. Katz, and D. Yuret. Word Sense Disambiguation for Information Retrieval. AAAI/IAAI 1999: 985.

[Kaufman et al, 1990] L. Kaufman, and P. J. Rousseeuw. Finding Groups in Data: An Introduction to ClusterAnalysis. John Wiley & Sons, 1990.

[Kaufmann, 1973] A. Kaufmann. Introduction à la théorie des sous ensembles flous, tome 1, Masson, 1973.

[Keen, 1992] E. M. Keen. Some aspects of proximity searching in text retrieval systems. Journal of Information Science, 18 :89.98,1992.

[Khan, 2000] L. R. Khan. Ontology-based Information Selection, Phd Thesis, Faculty of the Graduate School, University of Southern California. August 2000.

[Kleinberg, 1999] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM) 46(5), 604–632. 1999.

[Kleinberg et al, 1999] J. M. Kleinberg, R. Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew S. Tomkins. The Web as a Graph: Measurements, Models, and Methods. Book Series Lecture Notes in Computer Science. Éditeur Springer Berlin / Heidelberg. Volume 1627/1999. Pages 1-17. 1 janvier 1999.

[Kraft et al, 1983] D. H. Kraft, and D. A. Buell. Fuzzy Sets and Generalized Boolean Retrieval Systems. International Journal of Man-Machine Studies, 19(1): 45-56, 1983.

[Kraft et al, 1999] D. H. Kraft, G. Bordogna, and G. Pasi. Fuzzy set techniques in information retrieval, In Fuzzy Sets in Approximate Reasoning and Information Systems, (J. C. Bezdek, D. Dubois, H. Prade, eds.), Kluwer, 469-510, 1999.

[Krovetz et al, 1992] R. Krovetz, and W.B. Croft. Lexical ambiguity and information retrieval, ACM Trans. On Information Systems, 10(2), 115-141, 1992.

[Krovetz, 1997] R. Krovetz. (1997). Homonomy and Polysemy in Information Retrieval, in the Proceedings of the COLING/ACL '97 conference.

[Kuhn, 1960] M.E. Maron, and J.L. Kuhn. (1960). On relevance, probabilistic indexing, and information retrieval. Journal of the Association for Computing Machinery, 7(3), 216-244.

[Kupiec et al, 1995] J. Kupiec, J. Pedersen, et F. Chen. A trainable document summarizer. Dans les actes de 18th annual international ACM SIGIR conference on Research and

[Kwong, 2001] O. Y. Kwong. ``Word sense disambiguation with an integrated lexical resource." In: Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources, Pittsburgh, June 2001. http://www.seas.smu.edu/~rada/mwnw/papers/WNW-NAACL-206.pdf.gz

[Lakoff, 1973] G. Lakoff. Hedges: A study in meaning criteria and the logic of fuzzy concepts. J. of Philosophical Logic, 2, 458-508, 1973.

[Lakoff, 1987] G. Lakoff. Women, Fire and Dangerous Things. What Categories Reveal about the Mind. The University of Chicago Press, 1987.

[Larry et al, 1998] P. Larry, B. Sergey, and R. Motwani. (1998). "The PageRank citation ranking: Bringing order to the web," Online: http://citeseer. nj. nec.com/page98pagerank. html[04.06. 2003] chap2, 1998.

[Lawler et al, 1987] E.L. Lawler, J.K. Lenstra, A.H.G. Rinnooy Kan, D.B. Shmoys, J. Wiley and sons. The Traveling Salesman Problem: a Guided Tour of Combinatorial Optimization, New York, 1987 (ISBN 0-471-90413-9).

[Leacock et al, 1998] C. Leacock, G. A. Miller, and M. Chodorow. (1998). Using corpus statistics and WordNet relations for sense identification. Comput. Linguist. 24, 1 (Mar. 1998), 147-165.

[Lebart & Salem, 1994] L. Lebart, and A. Salem. Statistique textuelle. Paris : Dunod.

[Lin, 1999] C. Y. Lin, 1999. Training a selection function for extraction. Dans les actes de 8th international Conference on Information and Knowledge Management (CIKM), 55–62. ACM Press New York, NY, USA. 1999.

[Lin et al, 2004] T.Y. Lin, and I.J. Chiang. (2004). Automatic document clustering of concept hypergraph decompositions. In Proceedings of SPIE. Vol. 5098. (pp. 168-177). Orlando, FL.

[Lin, 1998] D. Lin. (1998) An information-theoretic definition of similarity. In Proceedings of 15th International Conference On Machine Learning, 1998.

[Lin et Hovy, 2003] C. Y. Lin et E. Hovy. The potential and limitations of automatic sentence extraction for summarization. Dans les actes de Human Language Technologies (HLT-NAACL), 73–80. Association for Computational Linguistics Morristown, NJ, USA. 2003.

[Losász et al, 1986] L. Lovász, and M.D. Plummer. Matching Theory, Annals of Discrete Mathematics 29, North-Holland, 1986 (ISBN 0-444-87916-1) - et aussi Akadémia Kiadó, Budapest, 1986.

[Losee, 1998] R.M. Losee. (1998). Text Retrieval and Filtering: Analytic Models of Performance. Kluwer, Boston.

[Luhn, 1957] H.P. Luhn. A Statistical Approach to Mechanized Encoding and Searching of Literary Information, IBM J. Research and Development, vol. 1, no. 4, pp. 309-317, 1957.

[Luhn, 1958] H. P. Luhn. The Automatic Creation of Literature Abstracts. IBM Journal of Research and Development 2(2), 159–165. 1958.

[Luhn, 1958]H. P. Luhn. The Automatic Creation of Literature Abstracts. IBM Journal of Research and Development 2(2), 159–165. 1958.

[Magnini, 2000] B. Magnini, and G. Cavaglià. (2000). Integrating Subject Field Codes into WordNet. Actes de LREC-2000, Second International Conference on Language Resources and Evaluation, Athènes, Grèce, pp. 1413-1418.

[Mani et al, 1999] I. Mani et M. T. Maybury. Advances in Automatic Text Summarization. The MIT Press. 1999.

[Mann et Thompson, 1988] W. C. Mann et S. A. Thompson. Rhetorical Structure Theory : A Theory of Text Organization. Text 8(3), 243–281. 1988.

[Marcu, 1997] D. Marcu. From discourse structures to text summaries. Dans les actes de ACL Workshop on Intelligent Scalable Text Summarization, 82–88. 1997.

[Maron & al, 1960] M.E. Maron, and J.L. Kuhn. “On relevance, probabilistic indexing, and information retrieval”. Journal of the Association for Computing Machinery, 7(3), 216-244. 1960.

[Mauldin, 1991] M.L. Mauldin. (1991). Retrieval performance in FERRET: a conceptual information retrieval system. In Proceedings of the lSth International A CM-SIGIR Conference on Research and Development in Information Retrieval,pages 347-355, Chicago, IL, October.

[Mihalcea et al, 2000] R. Mihalcea, and D. Moldovan. Semantic indexing using WordNet senses. In Proceedings of ACL Workshop on IR & NLP, Hong Kong, October 2000. http://www.seas.smu.edu/~rada/papers/acl00.nlp_ir.ps.gz

[Mihalcea, 2004] R. Mihalcea. Graph-based ranking algorithms for sentence extraction, applied to text summarization. Dans les actes de ACL 2004 on Interactive poster and

[Mihalcea et al, 2004] R. Mihalcea. Co-training and Selftraining for Word Sense Disambiguation. In In Proceedings of the Conference on Natural Language Learning (CoNLL 2004), Boston, USA, 2004.

[Miller et al, 1990] G. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. Introduction to WordNet: An on-line lexical database.J. of Lexicog.,3, 235-244, 1990.

[Miller, 1995] G. Miller. Wordnet: A lexical database. Communication of the ACM, 38(11):39--41, (1995).

[Minsky, 1975] M. Minsky. A Framework for Representing Knowledge. In The Psychology of Computer Vision, edited by P. Winston. New_York: MC-Graw_Hill, 1975. P 211-281.

[Mitra et al, 1997] M. Mitra, C. Buckley, A. Singhal, and C. Cardie. An analysis of statistical and syntactic phrases. In Proceedings of RIAO97, Computer--Assisted Information Searching on the internet, pages 200–214, Montreal, Canada, June 1997.

[Montes-y-Gómez et al, 2000] M. Montes-y-Gómez, A. López-López, and A. Gelbukh. Information retrieval with conceptual graph matching. Proc. DEXA'00, Greenwich, 2000. Springer, LNCS 1873, 312–321, 2000.

[Mooers, 1948] C.N. Mooers. Application of Random Codes to the Gathering of Statistical Information, MIT Master's Thesis, 1948.

[Mouton et al, 2009] C. Mouton, B. Richert, and G. Chalendar. Traduction de FrameNet par dictionnaires bilingues avec évaluation sur la paire anglais-français. Dans MajecSTIC 2009, Avignon, France, du 16 au 18 novembre 2009.

[Müller et al, 2004] H. M. Muller, E. E. Kenny, and P. W. Sternberg. Textpresso: An Ontology-Based Information Retrieval and Extraction System for Biological Literature. PLoS Biol. 2(11), E309, 2004.

[Mutschke, 2001] P. Mutschke. Enhancing Information Retrieval in Federated Bibliographic Data Sources Using Author Network Based Stratagems. SpringerLink Date lundi 1 janvier 2001. Book Research and Advanced Technology for Digital Libraries Pages 287-299.

[Morris et al, 2001] J. Morris, G. Hirst, Lexical cohesion computed by thesaural relations as an indicator of the structure of text, Computational Linguistics 17 (1) 21–43. 1991.

[Navigli et al, 2003] R. Navigli, and P. Velardi. An Analysis of Ontology-based Query Expansion Strategies. In 2003 Workshop on Adaptive Text Extraction and Mining held in conjunction with: 14th European Conference on Machine Learning (ECML). www.dsi.uniroma1.it/~navigli/ pubs/ECML_2003_Navigli_Velardi.pdf

[Negoita, 1973] C. V. Negoita. On the notion of relevance in information retrieval. Kybern. 2, 161-165, 1973.

[Newman, 2003] M.E.J. Newman. (2003). Arxiv preprint http://arxiv.org/abs/cond-mat/0309045.

[Ng, 1994] R.T. Ng, and J. Han. (1994). Efficient and effective clustering methods for spatial data mining. Proceedings of the 20th VLDB Conference, Santiago, Chile, pp. 144–155.

[Nie et al, 1999] F. Ren, L. Fan, and J.-Y. Nie. Approach: How to Acquire Knowledge in an Actual Application System, IASTED International Conference on Artificial Intelligence and Soft Computing, Honolulu, 1999, pp.136-140.

[Nieminen, 1974] J. Nieminen. “On the centrality in a graph,” Scandinavian Journal of Psychology, vol. 15, no. 1, pp. 332–336, 1974.

[Noh et al, 2004] J. D. Noh, and H. Rieger. Phys. Rev. Lett. 92, 118701 (2004).

[Ono et al, 1994] K. Ono, K. Sumita, et S. Miike. Abstract generation based on rhetorical structure extraction. Dans les actes de 15th conference on Computational linguistics, Volume 1, 344–348. Association for Computational Linguistics Morristown, NJ, USA. 1994.

[Opsahl et al, 2010] T. Opsahl, F. Agneessens, and J. Skvoretz. (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. Social Networks, doi: 10.1016/j.socnet.2010.03.006, http://toreopsahl.com/2010/04/21/article-node-centrality-in-weighted-networks-generalizing-degree-and-shortest-paths/

[Oueslati, 1999] R. Oueslati. (1999). Aide à l’acquisition de connaissances à partir de corpus. Rapport interne, Université Louis Pasteur Strasbourg. Thèse de Doctorat en Informatique.

[Perron, 1996] J. Perron. (1996). Adepte-Nomino : un outil de veille terminologique, dans Terminologies nouvelles, no 15, juin et décembre, Bruxelles, RINT, p. 32-47.

[Borland, 2003] P. Borlund. The concept of relevance in ir. J. Am. Soc. Inf. Sci. Technol., 54(10) :913–925, 2003.

[Pollock et Zamora, 1975] J. J. Pollock et A. Zamora. Automatic Abstracting Research at Chemical Abstracts Service. Journal of Chemical Information and Computer Sciences 15(4), 226–232. 1975.

[Ponte et Croft, 1998] J.M. Ponte, and W.B. Croft. A language modeling approach to information retrieval. In Proceedings of the 21st ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98), 1998, pp. 275 – 281.

[Porter, 1980] M.F. Porter. An algorithm for suffix stripping, Program, 14(13): 130-137, 1980.

[Qui et Frei, 1993] Y. Qui, and H. P. Feri. “Concept Based Query Expansion,” in Proc. of the Sixteenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, pp. 160-169, 1993.

[Quillian, 1968] M.R. Quillian. Semantic Memory. In Semantkic Informattion Processing, Edited by M. Minsky. Cambridge: MIT Press, 1968. P. 227-270.

[Radev et al, 2000] D. R. Radev, H. Jing, & M. Budzikowska. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In ANLP/NAACL Workshop on Summarization Seattle, WA. 2000.

[Radev et al, 2001] D. Radev, S. Blair-Goldensohn, & Z. Zhang. Experiments in single and multidocument summarization using MEAD. In First Document Understanding Conference New Orleans, LA. 2001.

[Rasolofo et Savoy, 2003] Y. Rasolofo, and J. Savoy. Term proximity scoring for keyword-based retrieval systems. In proceedings of the 25th European Conference on Information Retrieval, ECIR 2003 proceedings, number 2633 in LNCS, pages 207.218. Springer, 2003.

[Ren et al, 1999] F. Ren, L. Fan, and J.-Y. Nie Saak. Approach : How to Aquire Knowledge in an Actual Application System, IASTED International Conference on Artificial Intelligence Natural Language Processing, Washington, DC.

[Resnik, 1995] P. Resnik. "Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language", Journal of Artificial Intelligence Research (JAIR), 11, pp. 95-130, 1999.

[Resnik, 1999] P. Resnik. (1999). Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. Journal of Artificial Intelligence Research (JAIR), 11, 95-130.

[Rieger, 1978] B. B. Rieger. Feasible fuzzy semantics, Proc. 7th Int. Conf. on Computa. Linguis. (COLING 78), Bergen, (Heggstad, K., ed.), 41-43, 1978.

[Rieger, 2001] B. B. Rieger. Computing granular word meanings. A fuzzy linguistic approach to computational semiotics, In Computing with Words (P. P Wang, ed.), Wiley, 147-208, 2001.

[Van Rijsbergen, 1977] C. J. Van Rijsbergen. A theoretical basis for the use of co-occurrence data in information retrieval. Journal of Documentation, 33: 106-119, 1977.

[Van Rijsbergen, 1986] C. J. Van Rijsbergen. A new theorical framework for information retrieval. In Proceedings of the 9th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Pisa, Italy, Septembre 1986. P 194-200.

[Robertson et al, 1981] S. E. Robertson, C. J. van Rijsbergen, and M. F. Porter. (1981). Probabilistic models of indexing and searching. R.N. Oddy, S.E. Robertson, C.J. van Rijsbergen, and P.W. Williams, editors, Information Retrieval Research, (pp. 35-56). London: Butterworths.

[Robertson et al, 1976] S. E. Robertson, and K. Sparck Jones. (1976). Relevance weighting of search terms. Journal of the American Society for Information Science, 27, 129–146.

[Robertson et al, 1994] S. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3, NIST Special Publication 500-225: the Third Text REtrieval Conference (TREC-3), pp. 109-126.

[Robertson et al, 1997] S. E. Robertson, and S. Walker. On relevance weights with little relevance information. In Proceedings of the 20th annual international ACM SIGIR conference on Research and development in information retrieval, pages 16–24. ACM Press, 1997.

[Robertson, 1994a] S. Robertson, S. Walker, S. Jones, and M. H.-B. Gatford. « Okapi at 3 », Proceedings of the 3rd Text REtrieval Conference (-3), p. 109-126, 1994.

[Robertson, 1994b] S. E. Robertso, and S. Walker. « Some Simple Effective Approximations to the2-Poisson Model for Probabilistic Weighted Retrieval », Proceedings of SIGIR 1994, p. 232-241, 1994.

[Rocchio, 1971] J.I Rocchio, « Relevance feedback in information retrieval. In The SMART Retrieval System », dans Prentice-Hall, p. 313-323, 1971

[Roussey, 2001] C. Roussey. Une méthode d'indexation sémantique adaptée aux corpus multilingues, informatique, Lyon, these de l'INSA de Lyon, 2001, 150 pages.

[Roussey et al, 2001] C. Roussey, S. Calabretto, J-M. Pinon. Multinlinguak Information System based on Knowledge Representation. In Proceedings of the 5th International Conference on Advances in Databases and Information Systems (ADBIS'2001); Vilnius, Lithuania, 25-28 September 2001. p. 98-111. (Lectures Notes In Computer Science, Vol. 2151)

[Ruiz-Casado, 2005] M. Ruiz-Casado. Alfonseca E., Castells P. 2005. Automatic assignment of Wikipedia encyclopedic entries to WordNet synsets. Actes de AWIC, 380-386.

[Sabidussi, 1966] G. Sabidussi. "The centrality index of a graph," Psychometrika, vol. 31, no. 4, pp. 581–603, 1966.

[Salton et al, 1983] G. Salton, E.A. Fox, and H. Wu. Extended Boolean information retrieval system. CACM 26(11), pp. 1022-1036, 1983.

[Salton et al, 1968] G. Salton, and M. Lesk. Computer Evaluation of Indexing and Text Processing. J. ACM 15(1): 8-36 (1968).

[Salton et al, 1988] G. Salton, and C. Buckley. (1988). Term-weighting approaches in automatic text retrieval. Information Processing and Management: an International Journal 24 (5), 513-523.

[Salton et McGill, 1983] G. Salton, and M. McGill. Introduction to Modern Information Retrieval. McGraw-Hill, New York, 1983.

[Salton, 1971a] G. Salton. The SMART Retrieval System - Experiments in Automatic Document Processing. Prentice Hall, Englewood, Cliffs, New Jersey, 1971.

[Salton, 1971b] G. Salton. "A Comparison between manual and automatic indexing methods". Journal of the American Documentation, 20(1), pp. 6171, 1971.

[Salton, 1971c] G. Salton. The SMART Retrieval System: Experiments in Automatic Document Processing, 1971, Prentice-Hall, (Z699.4 S2 S25).

[Sanderson, 1994] M. Sanderson. Word sense disambiguation and information retrieval, Proc. of ACM SIGIR'94 Conf., 17, 142-151, 1994.

[Sanderson, 1997] M. Sanderson. Word Sense Disambiguation and Information Retrieval, PhD Thesis, Technical Report (TR-1997-7) of the Department of Computing Science at the University of Glasgow, Glasgow G12 8QQ, UK, 1997.

[Sarasevic, 70] T. Saracevic. “ he concept of “relevance” in information science : a historical review”, dans T Saracevic (dir), Introduction to Information science. R.R. Bowker, New York, p.111-151,1970.

[Svore et al, 2007] K. M. Svore, L. Vanderwende, et C. J. C. Burges. Enhancing Single-document Summarization by Combining RankNet and Third-party Sources. Dans les actes de Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), 448–457. 2007.

[Savoy, 1993] J. Savoy. Stemming of French words based on grammatical categories, Journal of the Americal Society for Information Science, 44(1): 1-9, 1993.

[Schmid, 1994] H. Schmid. Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing. Sept. 1994.

[Schuler, 2005] K. K. Schuler. VerbNet : A broad-coverage, comprehensive verb lexicon. Univ. of Pennsylvania-Electronic Dissertations, 2005.

[Scott et al, 1971] A. J. Scott, and M. J. Symons. (1971). clustering Methods based on Likelihood Ratio Criteria, Bometrics, 27, 387-397.

[Sheikholeslami, 1998] G. Sheikholeslami, S. Chatterjee, and A. Zhang. WaveCluster: A Multi-Resolution Clustering Approach for Very Large Spatial Databases. In Proceedings of the 24th International Conference on Very Large Data Bases, (VLDB), pages 428–439, New York, NY, USA, 24–27 August 1998. Morgan Kaufmann Publishers.

[Sheridan et al, 1992] P. Sheridan and A. F. Smeaton. The application of morpho-syntactic language processing to effective phrase matching. Inf. Process. Manage., 28(3) :349–370, 1992.

[Singhal et al, 1996] A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In Proceedings of the 19th Annual International ACM SIGIR Conference on Researchand Development in Information Retrieval, pages 21–29, 1996.

[Singhal et al, 1997] A. Singhal, M. Mitra, and C. Buckley. (1997). Learning routing queries in a query zone. In Proceedings of the 20th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Philadelphia, Pennsylvania, United States, July 27 - 31, 1997). N. J. Belkin, A. D. Narasimhalu, P. Willett, and W. Hersh, Eds. SIGIR '97. ACM Press, New York, NY, 25-32.

[Smadja, 1993] F. Smadja. Retrieving collocations from text : Xtract. Computational Linguistics, 19(1), pp: 143-177.

[Sowa, 1984] J. Sowa. Conceptual Structures: information processing in mind and machine. In The System Programming Series, Reading: Addision Wesley publishing Company, 1984. 481 pages.

[SparckJones, 1979] K. Sparck Jones. Experiments in relevance weighting of search terms. Inf. Process. Manage. 15(3): 133-144, 1979.

[SparckJones, 1986] K. Sparck Jones. Issues in User Modelling for Expert Systems. In Artificial Intelligence and its Applications, A. G. Cohn and J. R. Thomas edit., Wiley. pp. 183-195.

[Späth, 1985] H. Späth. Cluster dissection and Analysis, (Ellis Horwood, Chichester).

[Stairmand et al, 1996] M. A. Stairmand, and J. B. William. (1996). "Conceptual and Contextual Indexing of Documents using WordNet-derived Lexical Chains." In Proc. of 18th BCS-IRSG Annual Colloquium on Information Retrieval Research.

[Stein et al, 1997] A. Stein, J.A. Gulla, A. Müller, and U. Thiel. (1997). Conversational interaction for semantic access to multimedia information. In Maybury, M. T., ed., IntelligentMultimedia Information Retrieval.Menlo Park, CA: AAAI/The MIT Press. 399–421.

[Stephenson et al, 1989] K.A. Stephenson, and M. Zelen. (1989). Rethinking centrality: Methods and examples. Social Networks 11, 1–37.

[St-Jacques et al, 2004] C. St-Jacques, C. Barrière, and H. Prade. Fuzzy logic tools for lexical acquisition. Proc. 10th IPMU'04, Perugia, July 4-9, 2004, 2045-2052.

[Strapparava et al,2004] C. Strapparava, and A. Valitutti. (2004). WordNet-Affect: an Affective Extension of WordNet. Actes de 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisbonne, pp. 1083-1086.

[Strzalkowski et al, 1998] Tomek Strzalkowski, Gees C. Stein, G. Bowden Wise, Jose Perez Carballo, Pasi Tapanainen, Timo Jarvinen, tro outilainen, and Jussi Karlgren. Natural language information retrieval : TREC-7 report. In Text REtrieval Conference, pages 164–173, 1998.

[Subasic et al, 2000] P. Subasic, and A. Huettner. Calculus of fuzzy semantic typing for qualitative analysis of text, ACM KDD 2000 Workshop on Text Mining, Boston, 2000.

[Symons, 1981] M. J. Symons. (1981), Clustering Criteria and Multivariate Normal Mixture, Biometrics, 37. 35-43.

[Thomopoulos et al, 2003] R. Thomopoulos, P. Buch, and O. Haemmerlé. Representation of weakly structured imprecise data for fuzzy requiting. Fuzzy Sets and Systems, 140, 111-128, 2003.

[Vaina, 1978] L. M. Vaina. Semiotics of with, Versus, 17, (Bompiani, Milano), 96-112, 1978.

[Vaina, 1980] L. M. Vaina. Fuzzy sets in the semiotic of text, Semiotica, 31, 261-272, 1980.

[vallette, 2009] M. Valette. Approche textuelle du lexique. Mémoire habilitation à diriger des recherches. Institut National des Langues et Civilisations Orientales. Novembre 2009.

[van Rijsbergen, 1979] C. J. van Rijsbergen. Information Retrieval. Butterworth 1979.

[Voorhees, 1986] E. M. Voorhees. Implementing Agglomerative Hierarchical Clustering Algorithms for use in document retrieval, Information Processing & Management, 22 (6) 465-476, (1986).

[Voorhees, 1993] E. Voorhees. "Using WordNet to Disambiguate Word Senses for Text Retrieval", Proceedings of the 16thAnnual Conference on Research and Development in Information Retrieval, SIGIR'93, Pittsburgh, PA, 1993.

[Warshall, 1962] S. Warshall. “A theorem on boolean matrices,” J. ACM, vol. 9, no. 1, pp. 11–12, 1962.

[Wasserman et al, 1994] S. Wasserman, and K. Faust. Social Network Analysis: Methods and Applications. Cambridge University Press, 1994.

[WeiWang et al, 1997] J. Y. WeiWang, and R. M. Richard. STING:AStatistical Information Grid Approach to Spatial Data Mining. In Proceedings of the 23rd International Conference onVery Large Data Bases, (VLDB), pages 186–195, Athens, Greece, 26–29 August 1997. Morgan Kaufmann Publishers.

[Wong et al, 1985] S.K.M. Wong, W. Ziarko and P.C. N. Wong : Generalized Vector Space Model in Information Retrieval. In Proc of the 8th ACM SIGIR Conference on Research and Development, New-York USA, 1985

[Woods, 1997] W.A. Woods. Conceptual indexing : A better way to organize knowledge. Technical Report SMLI TR-97-61, Sun Microsystems Laboratories. P.18, Mountain View, CA, April, 1997. (available online at: http://research.sun.com/techrep/1997/abstract-61.html).

[Yang, 1999] Y. Yang. An evaluation of statistical approaches to text categorization. Information Retrieval, vol. 1(1-2): p. 69-90, 1999.

[Yarowsky, 1995] D. Yarowsky. «Unsupervised word sense disambiguation rivaling supervised methods», In 33rd Annual Meeting, Association for Computational Linguistics, Cambridge, Massachusetts, USA, 1995, (p189-196).

[zadeh 1965] L. A. Zadeh. Fuzzy sets, Information and Control, 8, 338-353, 1965.

[Zadeh, 1971] L. A. Zadeh. Quantitative fuzzy semantics, Information Sciences, 3, 159-176, 1971.

[Zhai, 2009] Ch. Zhai. Statistical Language Models for Information Retrieval. Morgan & Claypool, 2009.

[Zhang et al, 1996] T. Zhang, R. Ramakrishnan, and M. Livny. (1996). BIRCH: An efficient data clustering method for very large databases. Proceedings of ACM SIGMOD Conference, Montreal, Canada, pp. 103–114.

[Zhuge et al, 2007] H. Zhuge and X. Li. “Peer-to-Peer in Metric Space and Semantic Space,” IEEE Transactions on Knowledge and Data Engineering, vol. 6, no. 19, pp. 759–771, 2007.

[Zhuge et al, 2009] H. Zhuge, J. Zhang. Topological Centrality and Its Applications. Cite as: arXiv:0902.1911v1 [cs.IR] chap2 11 Feb 2009.

[Zipf, 1949] G.K. Zipf. Human Behavior and the Principale of Least Effort, Addision-Welsey, Cambridge, MA. 1949.

PARTIE III : **Annexes**

1. Autres ressources pour l'extraction des descirpteurs
2. La centralité
3. Tableaux et résultats

1. Autres ressources linguistiques

Des ressources linguistiques autres que WordNet existent pour représneter les concepts d'un document. Ils prennent différentes formes décrivons quelques-unes :

VerbNet[14] : VerbNet est une partie de projet SemLink en cours de développement à l'université de Colorado qui est un prolongement de travaux de Levin [Levin, 2003]. C'est un lexique des classes de verbes anglais [Schuler, 2005] [Chaumartin, 2007b], dont chaque classe regroupe des verbes partageant les mêmes comportements syntaxiques et sémantiques.

La version VerbNet 2.1 distingue 237 classes de verbes qui regroupent 4991 sens de verbes. Chaque verbe membre d'une classe il est accompagné de *synset* qui lui correspond dans WordNet pour définir le sens précis. En plus, VerbNet dispose d'un mappage vers FrameNet et une API en Java.

FrameNet : FrameNet est un projet mené à l'*International Computer Science Institue* de Berkeley, California qui est fondé sur la sémantique des cadres (« *semantic frames* ») [Fillmore et al, 2002] inspiré des travaux de Fillmore en 1968[Fillmore, 1968]. Dans FrameNet, les unités lexicales sont livrées accompagnées d'une notice d'actualisation décrivant la combinaison syntaxique (arguments) et sémantique (actants) de leurs différentes acceptions [Vallette, 2009].

Figure 18: Exemple de relations entres cadres

FramNet définit un certain nombre de cadres appelés *Frames* décrivant chacun une situation précise susceptible d'apparaître dans notre perception du monde. Les rôles impliqués dans cette situation sont répertoriés comme dans la Figure 18, ainsi que, les unités lexicales (*Lexical Unit, LU*) déclencheuses de cette situation, c'est-à-dire les prédicats (verbes, noms, adjectifs, ou mêmes adverbes) régisseurs de ces rôles [Mouton et al, 2009].

EXtended WordNet[15] **:** eXtended WordNet (XWN) est un projet développé à *Human Language Technology Research Institute* de l'université de Texas à Dallas (et financé par *National Science Foundation*). Le but de ce projet consiste à développer un outil prenant comme entrée l'actuelle ou les futures versions de WordNet, ensuite, générée automatiquement une extension de WordNet (eXtended WordNet). Cette extension fournit des importantes améliorations dans le but de remédier aux limitations actuelles de WordNet.

XWN produit une analyse syntaxique de la définition de chaque synset, une désambigüisation lexicale de chaque mot de la définition, puis un passage en forme logique. XWN est accessible gratuitement sous la licence *BSD style License* (Berkeley Software Distribution style License). La dernière version n'a pas été modifiée depuis Novembre 2004 basée sur la version 2.0 de WordNet).

14 http://verbs.colorado.edu/verb_index
15 http://xwn.hlt.utdallas.edu/

WordNet Domains[16] : WordNet Domains est une ressource lexicale crée d'une manière semi-automatique pour augmenter WordNet avec des étiquettes de domaines [Magnini, 2000]. Chaque synset de WordNet est annoté par au moins une étiquette sémantique de domaines (par exemple Sport, Politique, Médecine, Economie..), choisit dans un ensemble d'environ de deux cents étiquettes organisés hiérarchiquement.

L'utilisation des domaines permet de compléter les informations présentées dans WordNet. Les domaines offrent une manière d'établir des relations sémantiques entre les sens des mots, qui peuvent être utilisés avec profit en informatique linguistique. Un domaine peut inclure des synsets de différentes parties du discours et de différentes sous-hiérarchies de WordNet. Dans cette nouvelle structure, les différents sens d'un terme sont regroupés dans différents clusters homogènes, le fait qui réduit la polysémie des termes.

WordNet Domains a été intégré à la base lexicale multi-langue de WordNet *MultiWordNet*. Elle est distribuée pour des utilisations au niveau des recherches et aux niveaux commerciales.

WordNet-Affect : WordNet-Affect est une ressource linguistique pour la représentation lexicale de connaissances sur les affects [Strapparava et al, 2004]. Basée sur WordNet Domain, WordNet-Affect est une hiérarchie des concepts affectifs où les concepts sont des étiquettes de domaines composés des sous-ensembles de synsets de WordNet. On ajoute des informations additionnelles aux synsets affectifs, en leur associant une ou plusieurs étiquettes qui précisent une signification affective (exemple « joy », « Happy », « sad »,… qui marque une sensation positive, négative, neutre ou désambiguïsée).

Dans le Tableau 17, nous présentons la liste de l'étiquette affective, avec des exemples des synsets associés.

Etiquette affective	**Exemples de synsets associés**
Emotion	nom ANGER#1, verbe FEAR#1
Mood	nom ANIMOSITY#1, adjectif AMIABLE#1
Trait	nom AGGRESSIVENESS#1, adjectif COMPETITIVE#1
Cognitive State	nom CONFUSION#2, adjectif DAZED#2
Edonic Signal	nom HURT#3, nom SUFFERING#4
Emotion-Eliciting Situation	nom AWKWARDNESS#3, adjectif OUT OF DANGER#1
Emotional Response	nom COLD SWEAT#1, verbe TREMBLE#2
Behaviour	nom OFFENSE#1, adjectif INHIBITED#1
Attitude	nom INTOLERANCE#1, nom DEFENSIVE#1
Sensation	nom COLDNESS#1, verbe FEEL#3

Tableau 17: Liste des étiquettes effectives

WordNet-Affect est utilisée dans des intérêts économiques réels pour la détection de sensations affectives, par exemple, une société peut chercher dans la blogosphère ou les news, s'il se dit du bien ou du mal de ses produits.

16 http://wndomains.itc.it/

SentiWordNet[17] : SentiWordNet est une ressource lexicale pour la détection d'opinions [Esuli et al, 2006]. Elle assigne à chaque synset de WordNet 2.0 trois valeurs : Positivité, Négativité, Objectivité (respectant l'égalité : Positivité + Négativité + Objectivité = 1). Cette ressource a été créée d'une façon semi-automatisées, en mixant des techniques linguistiques et statistiques (utilisation de classifieurs) [Chaumartin, 2007b].

Dans Tableau 18 Exemple d'utilisation pour les 3 sens de l'adjectif « estimable ».

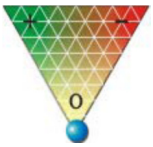
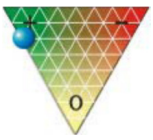
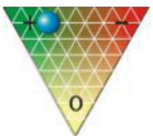
	P = 0 N = 0 O = 1	COMPUTABLE#1 **ESTIMABLE**#3 *may be computed or estimated; "a calculable risk"; "computable odds"; "estimable assets"*
	P = 0,75 N = 0 O = 0,25	**ESTIMABLE**#1 *deserving of respect or high regard*
	P = 0,625 N = 0,25 O = 0.125	HONORABLE#5 GOOD#4 RESPECTABLE#2 **ESTIMABLE**#2 *deserving of esteem and respect; "all respectable companies give guarantees"; "ruined the family's good name"*

Tableau 18: Degré d'opinion affecté pour les 3 sens de l'adjectif "estimable" [Chaumartin, 2007b]

Wikipédia : Wikipédia (prononcé /wi.ki.pe.dja/) est une encyclopédie, multilingue, universelle, librement diffusable, disponible sur le web et écrite par les internautes grâce à la technologie wiki. Elle a été créée en janvier 2001 et est devenue un des sites web les plus consultés au monde. Elle est hébergée par une fondation américaine, la Wikimedia Foundation. Plusieurs projets visent à établir automatiquement des liens entre la Wikipédia et WordNet.

Certains algorithmes réalisent une correspondance entre un article de la version anglaise de Wikipedia et le synset correspondant de WordNet [Ruiz-Casado, 2005]. Suivant les auteurs, ce type des implémentations revendique une précision de 91,11% (83,89% sur les mots polysémiques). [Chaumartin, 2007a] propose une généralisation de ce type d'approche, où WordNet est en plus enrichi avec des nouveaux synsets, avec une identification du bon hyperonyme. La précision de l'appariement entre WordNet 2.1 et un sous-ensemble de la Wikipedia anglaise (autour de 15 800 articles) est de 92% ; en cas de création de nouveau synset, l'hyperonyme est correctement identifié dans 85% des cas.

Le projet DBpedia[18], extrait des informations structurées à partir de Wikipedia et les rend disponible sur le Web. Le projet consiste à convertir le contenu de Wikipédia à des données structurées en utilisant les techniques du Web sémantiques afin de, pouvoir exécuter des requêtes sophistiquées sur Wikipédia, lier son contenue à d'autres ensembles de données sur le Web. En plus, DBpédia couvre de nombreux domaines, ce qui n'est pas le cas de la plupart des bases de connaissances qui couvrent que des domaines spécifiques.

La base de connaissance de DBpédia est constituée d'environ 2,6 millions d'objets, incluant 213000 personnes, 328000 lieux, 57000 albums musicaux, 36000 films, 20000 entreprises, sauvegardés chacun sous forme de document RDF[19], dans la Figure 19 nous

17 http://sentiwordnet.isti.cnr.it/
18 http://wiki.dbpedia.org
19 http://www.w3.org/RDF

présentons un extrait de la hiérarchie de DBpédia. Une part importante de ces ressources est disponible dans 36 langues différentes sachant qu'elles évoluent automatiquement avec les changements de Wikipedia.

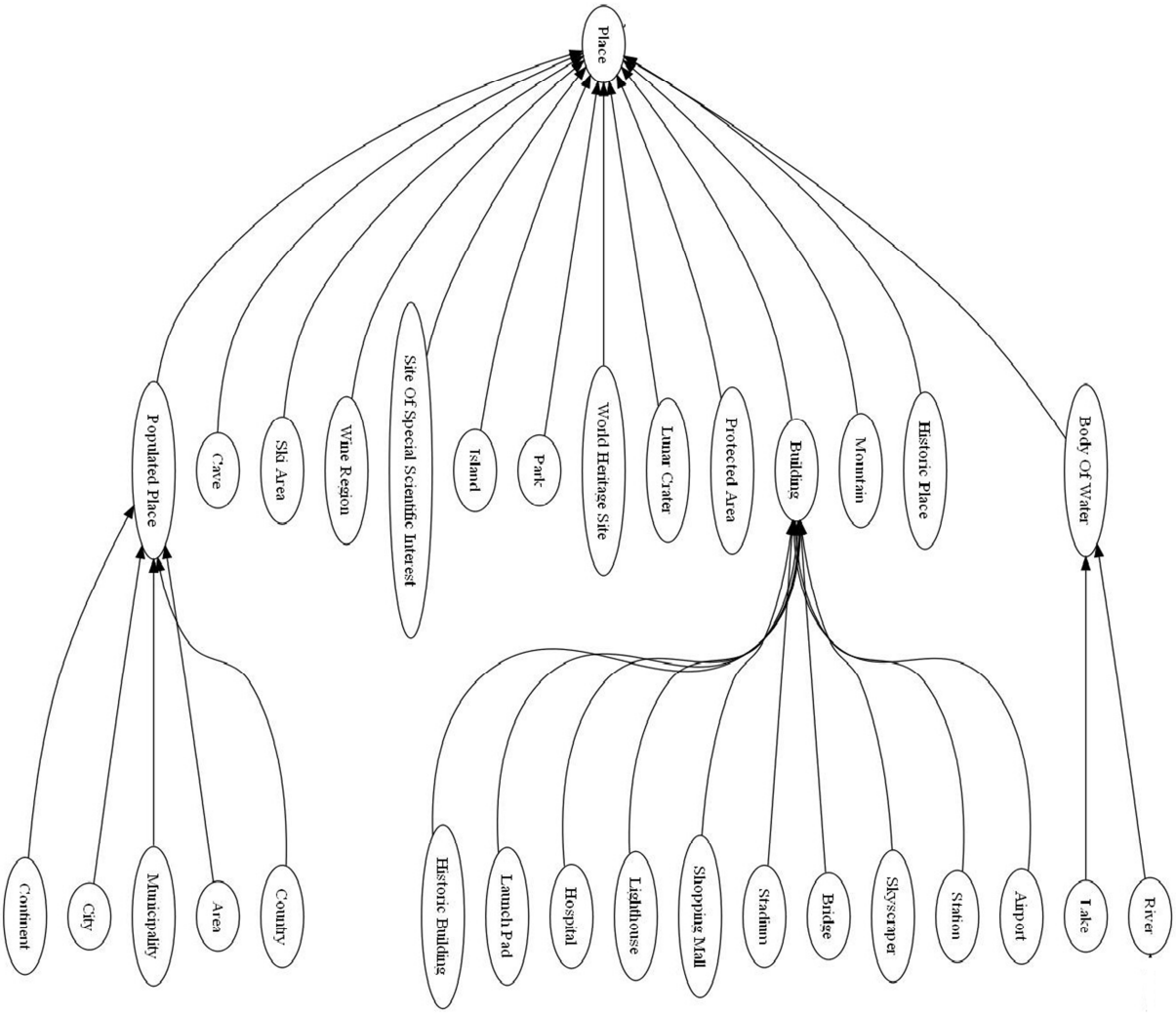


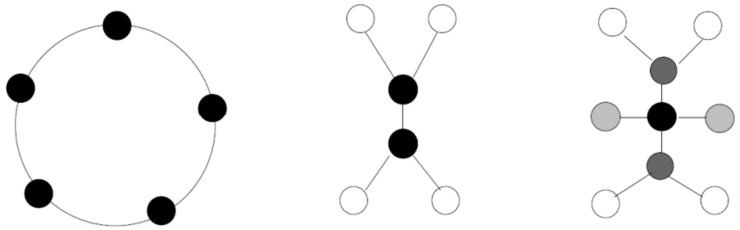
Figure 19: Extrait de la base de connaissance DBpédia

2. La centralité

L'idée de la centralité est reliée à la notion des graphes. Les différentes mesures de centralité d'un nœud (ou *sommet*) dans un graphe correspondent à déterminer son importance dans ce graphe (par exemple, le calcul du degré d'importance d'une personne dans un réseau social ou autre exemple, quel est le score d'utilisation d'une route dans un réseau d'urbanisme ?). Selon les observations de Zhuge et al. [Zhuge et al, 2009], un nœud est plus important dans un graphe en deux cas : s'il est connecté à plusieurs nœuds et s'il est connecté à plusieurs nœuds importants.

La centralité a été traitée dans différents domaines : i) Dans le domaine des recherches sur le web pour développer des algorithmes de classifications des topiques, où les pages et les hyperliens vers les pages Web sont représentés comme des nœuds et des arêtes dans un graphe [Kleinberg et al, 1999] ; ii) Dans les réseaux sociaux, où la centralité est utilisée pour représenter l'acteur principal qui travaille sur un sujet central [Mutschke, 2001]. iii) Dans d'autres cas, la centralité est utilisée pour compter un degré de relation entre une entrée sémantique et les différents catégories [Huang et al, 2003]. iv) Dans le domaine des résumés textuels pour mesurer la centralité d'une phrase dans un document [Erkan et Radev, 2004], une méthode basée sur les graphes a été utilisée dans le processus de traitement de la langue pour mesurer l'importance d'une unité textuelle. Dans cette méthode, la centralité d'une phrase est définie par la centralité de ces termes. La méthode principale pour mesurer la centralité d'un terme c'est de chercher le centroïde d'un document dans un espace vectoriel. Le centroïde d'un document est une partie du document qui est constitué des termes qui ont un *tf*idf* supérieur à un degré prédéfini, avec *tf* est la fréquence d'un terme dans un cluster, et *idf* est une valeur calculée d'une manière similaire à celle calculer dans une collection de documents. Dans notre proposition, nous utilisons la notion de la centralité d'une manière différente, elle est utilisée comme un paramètre pour ordonner les documents pertinents qui correspondent à une requête dans le domaine de la recherche d'information.

Différents topologies peuvent exister pour la centralité. Dans [Zhuge et al, 2009], trois différentes topologies de la centralité ont été présentées qui dépendent de la structure du graphe : 1- un réseau avec une structure circulaire contenant n centres ($n \geq 3$) voir Figure 20 (a). 2- un réseau qui a une structure symétrique ayant deux centres topologiques voir Figure 20 (b). 3- un réseau qui a une structure générale ayant un centre topologique unique voir Figure 20 (c). Dans ces trois types de topologie, le nœud qui a la couleur la plus sombre représente le centre de la topologie.



(a) Structure circulaire (b) Structure symétrique (c) Structure générale

Figure 20: Trois structures topologiques

Les mesures de centralité ont été largement étudiées. Citons les mesures de centralité qui sont largement utilisées dans l'analyse des graphes : centralité de degré, centralité d'intermédiarité, centralité de proximité, centralité de prestige et centralité de pouvoir. C'est à l'aide de ces mesures que les chercheurs calculent les degrés de forces entre les nœuds d'un graphes [Opsahl et al, 2010]

Centralité de degré : elle décrit le degré d'information de chaque nœud [Freeman, 1979] [Nieminen, 1974]. Elle est basée sur le principe qu'un nœud est important dans un graphe s'il a plus des voisins. Cette mesure est utilisée pour trouver le nœud principal d'un graphe. la centralité de degré d'un sommet *v* dans un graphe *G* noté par *deg(v)* est estimée par la somme des liens existants avec les autres membres de graphes. Dans un graphe orienté, nous définissons deux mesures séparées pour la centralité : degré-entrant (*indegree* en anglais) et le degré-sortant (*outdegree* en anglais) [Zhuge et al, 2007]. Le degré-entrant est estimé par le nombre des liens directs vers le nœud, et le degré-sortant est estimé par le nombre des liens sortant du nœud dans la direction d'autres nœuds. Dans les réseaux sociaux par exemple, nous interprétons le degré-entrant pour estimer la popularité d'un acteur, et nous interprétons le degré-sortant pour estimer la sociabilité d'un acteur.

Dans un graphe *G = (V, E)* de *n* nœuds et *m* arêtes, le degré de centralité $C_D(v)$ d'un nœud (sommet) *v* est définie par :

$$C_D(v) = \frac{\deg(v)}{n-1}$$

Centralité d'intermédiarité : elle estime le degré auquel est lié un nœud aux autres nœuds dans un graphe qui ne sont pas forcément connecté à lui par des liens directes [Freeman, 1979] [Anthonisse, 1971][Freeman, 1977]. En d'autres termes, elle mesure si un nœud forme un nœud intermédiaire dans un graphe. L'idée est la suivante, si plusieurs nœuds sont connectés à travers un nœud, alors le nœud devient plus important.

Dans un graphe *G = (V,E)* de *n* nœuds et *m* arêtes, le degré de centralité $C_B(v)$ d'un nœud (sommet) *v* est définie par :

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

Où σ_{st} est estimé par le nombre de plus courts chemins de s vers t, et $\sigma_{st}(v)$ est estimé par le plus courts chemins de *s* vers *t* passant par v. cette formule est normalisée en divisons par le nombre des sommets (nœuds) ne contenant pas *v*, estimée par *(n-1)(n-2)* pour les graphes directes et *(n-1)(n-2)/2* pour les graphes indirectes.

Le chemin le plus court entre deux nœuds d'un graphe peut être trouvé par l'algorithme de Floyd-Warshall [Warshall, 1962] qui a une complexité en ordre de temps de $O(n^3)$. Sur les graphes non pondérés, le calcul de la centralité intermédiaire est d'ordre *O(nm)* de complexité de temps en utilisant l'algorithme de Brande [Brandes, 2001]. La centralité d'intermédiarité a été utilisée pour étudier la structure d'une communauté dans un réseau social et biologique [Girvan et al, 2002].

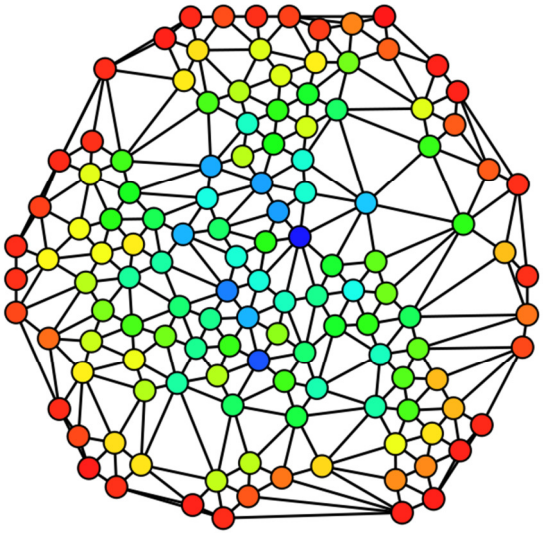


Figure 21: Exemple sur les noeuds intermediaire

Dans la Figure 21 nous donnons un exemple sur les nœuds intermédiaires dans un graphe. Le degré d'intermédiaire passe de couleur rouge pour indiquer un degré =0 au couleur bleu pour indiquer la plus grande centralité d'un nœud dans le graphe.

Centralité de proximité : cette centralité mesure le degré auquel un nœud est près de tous les autres individus dans un graphe (d'un façon directe ou non) [Freeman, 1979][Sabidussi, 1966] [Wasserman et al, 1994]. Elle permet d'identifier les nœuds qui ont la plus courte distance géodésique[20] par rapport aux autres nœuds du graphe dans le but d'identifier les nœuds sources d'un graphe. La centralité de proximité peut être considérée comme une mesure pour estimer la longueur qui peut prendre un nœud pour atteindre un autre nœud dans un graphe [Newman, 2003].

Dans la théorie des graphes, la centralité de proximité est une mesure sophistiquée de la centralité. Elle est définie comme la distance géodésique entre un nœud v et tous les autres nœuds accessibles à partir de ce nœud :

$$C_p(v) = \frac{\sum_{t \in V/v} d_G(v,t)}{n-1}$$

Où n ≥ 2 qui est le nombre des nœuds de V qui sont accessibles à partir de v.

Dans d'autres formules, la proximité est l'inverse de la distance géodésique entre chaque nœud du graphe [Sabidussi, 1966] :

$$C_p(v) = \frac{n-1}{\sum_{t \in V/v} d_G(v,t)}$$

Différentes méthodes et algorithmes sont introduites pour mesurer la proximité, exemple *random-walk centrality* introduites par Rieger [Noh et al, 2004], *information centrality* de Stephenson et Zelen [Stephenson et al, 1989], [Dangalchev, 2006], etc.

Centralité de prestige : décrit l'importance d'un nœud selon la matrice d'adjacence [Bonacich, 1972]. Cette mesure assigne des scores relatifs à tous les nœuds du graphe en se

20 En géométrie, une géodésique désigne le chemin le plus court, ou l'un des plus courts chemins s'il en existe plusieurs entre deux points d'un espace pourvu d'une métrique (un moyen de mesurer les distances)

basant sur le principe que les connexions vers les nœuds ayant les scores les plus élevés, contribuent davantage au score du nœud en question que des connexions égales mais à de plus bas score. Page Rank est une variante de mesure de la centralité de prestige [Larry et al, 1998].

Soit x_i signifie le score de du $i^{ème}$ nœud. Soit $A_{i,j}$ la matrice d'adjacence du graphe, d'où $A_{i,j} = 1$ si le $i^{ème}$ nœud est adjacent au $j^{ème}$ nœud, $A_{i,j} = 0$ dans les autres cas. D'une manière générale, les entrées dans la matrice *A* peuvent être des nombres réels qui représentent la force de connexion, comme le cas dans les matrices stochastiques21.

Pour le $i^{ème}$ nœud, le score de centralité est une proportion à la somme des scores de tous les nœuds qui lui sont connectés:

$$x_i = \frac{1}{\lambda} \sum_{j \in M(i)} x_j = \frac{1}{\lambda} \sum_{j=1}^{N} A_{i,j} x_j$$

où *M(i)* est l'ensemble des nœuds qui sont connectés au $i^{ème}$ nœud, *N* est le nombre total des nœuds et λ est une constante.

21 En mathématiques, une matrice stochastique (aussi appelée matrice de Markov) est une matrice carrée dont chaque élément est un réel compris entre 0 et 1 et dont la somme des éléments de chaque ligne vaut 1.

3. Tableaux et résultats

3.1. Vue sémantique des documents

Nous présentons dans le Tableau 19 la vue sémantique de la collection des documents d'Erika extrait en utilisant la fonction « $\min(\mu_{TF\text{-}grandsuffisamment}, \max(\mu_{s\text{-}grand},\mu_{c\text{-}grand}))$ » que nous utilisons pour extraire les phrases significatifs dans la section 4.4.2.

Doc. n°	Results with the function : $\min(\mu_{TF\text{-}grandsuffisamment}, \max(\mu_{s\text{-}grand},\mu_{c\text{-}grand}))$
1	1:{C6:{trial} C11:{company} C15:{tanker} C22:{sea} } 2:{C6:{plaintiff} C15:{ship} C56:{total} } 3:{C6:{court} C9:{million} C15:{oil} C19:{accident} C30:{disaster} C31:{damage} C37:{pollution} } 4:{C4:{responsibility, impunity} C6:{appeal, case} C7:{owner} C9:{ten_thousand, hundred} C10:{world} C11:{industry} C13:{lawyer} C15:{structure} C21:{december, february} C22:{coast} C34:{warning} C48:{dollar} C55:{eleven, manager} } 5:{C6:{defendant} C11:{organisation} C15:{fuel} C22:{france, bay} } 6:{C11:{party} C22:{shoreline, state} }
2	1:{C10:{ship} C18:{france, parisian, sea} } 2:{C10:{oil} C14:{court} C23:{oil_company} } 3:{C5:{uria_aalge} C10:{tanker, tank_ship} C11:{biodiversity} C13:{president} C29:{fratercula_arctica, atlantic_puffin} C30:{disaster} C32:{storm} C36:{kittiwake} C42:{sinking} C47:{spill} } 4:{C18:{brittany, coast} }
3	1:{C1:{seabird} C53:{oil} } 2:{C0:{france} C1:{bird} } 3:{C5:{tank} C17:{litre} C25:{damage} C29:{tanker} C32:{disaster, catastrophe} C48:{coastline} C51:{spill} }
4	1:{C6:{accusation} } 2:{C6:{charge} } 3:{C6:{pollution} } 4:{C5:{tanker} C6:{safety} C17:{company} } 5:{C10:{employee} C12:{accident} C18:{expert} C19:{investigation} C27:{probe} C36:{oil} C37:{judge} } 6:{C6:{condition} } 7:{C6:{statement, allegation, complicity} } 8:{C6:{authority, action} C9:{france, brittany} }
5	1:{C15:{mussel} } 2:{C1:{cat} C5:{ache} C15:{mytilus_edulis} C23:{year} C25:{survey} C29:{spill} C30:{oil} }
6	1:{C1:{court, tribunal} } 2:{C1:{case} C5:{oil} C9:{victory} C11:{plaintiff} C15:{damage} C19:{decision} C21:{spill} }
7	1:{C13:{matter, compound} } 2:{C13:{water, fuel, sediment} } 3:{C13:{coast, sulfur, oil} } 4:{C13:{shoreline} } 5:{C21:{contamination} C23:{mollusc} }
8	1:{C5:{trial, case} C21:{france, sea} } 2:{C20:{vote, decision} C21:{euro} C47:{total} } 3:{C5:{court} C11:{lawyer} C17:{accident} C19:{km} C31:{damage} C46:{oil} } 4:{C0:{norm, moment} C3:{january, december, february} C4:{tanker} C9:{industry, oil_company} C10:{tourism} C13:{profit} C21:{coast, bay} C25:{polluter, pollution} C33:{story} C36:{seabird} } 5:{C4:{ship, vessel} C21:{beach, shoreline, new_york} }
9	1:{C2:{france} } 2:{C2:{lawyer} } 3:{C2:{sea, state, euro} } 4:{C2:{atlantic, europe, president, day, january, december, february, christmas} } 5:{C2:{brittany} } 6:{C2:{paris, poitou-charentes, coast} C12:{trial, judge, ruling} } 7:{C1:{fuel_oil} C2:{bay, pays_de_la_loire, italy, shore} C25:{tanker} } 8:{C1:{substance} C12:{court} } 9:{C31:{damage} }
10	1:{C1:{oil, tanker} C2:{france, atlantic} C4:{amendment} C8:{tonne} C10:{april, december} C13:{disaster} } 2:{C2:{brittany, sea-coast} }
11	1:{C0:{scandal} C1:{reporter, tom} C24:{mangold} } 2:{C1:{individual} C6:{france, sea} C11:{oil_tanker, tanker} C15:{monday} }

	3:{C1:{official, agency} C6:{range} C11:{ship, oil} C15:{september, december} }
12	1:{C2:{eu} } 2:{C2:{country, state} } 3:{C0:{flag} C3:{market} C11:{parliament} C14:{damage} C25:{council} C26:{ship} C27:{oil} C28:{directive} C39:{standard} C45:{insurance} C47:{disaster} C48:{inspection} C60:{coastline} } 4:{C2:{france} C10:{polluter, pollution} C14:{incentive} C18:{obligation, enforcement} C36:{commission, authority} } 5:{C2:{port, sea, britain} }
13	1:{C0:{france, sea} } 2:{C0:{coast} C7:{trial} } 3:{C7:{case} } 4:{C7:{defendant, official} }
14	1:{C3:{judge} } 2:{C2:{trial, case} C3:{manager} C11:{tanker} C22:{damage} } 3:{C2:{court} C3:{party, association} C11:{oil} C14:{euro} } 4:{C18:{company} C19:{plaintiff} C22:{charge, cost} C44:{firm} C60:{disaster} C68:{spill} C69:{total} }
15	1:{C3:{ship} } 2:{C3:{trial, case} } 3:{C3:{charge, damage, tanker} } 4:{C3:{government, company, oil} } 5:{C3:{proceeding, tribunal, faith, individual, contract, article, witness, official, council, friend, certification, crew_member, lawyer, critic, law} } 6:{C0:{france} } 7:{C0:{sea} } 8:{C0:{euro} C3:{group} } 9:{C6:{feb} C14:{plaintiff} C21:{pollution} C61:{spill} C62:{total} } 10:{C0:{bay} C3:{organisation} C6:{june} }
16	1:{C1:{france, sea} } 2:{C1:{atlantic, bay, europe, italy, state, euro, president, day, december, week} } 3:{C1:{coast} C22:{tanker} C28:{damage} C36:{oil} C40:{bird} } 4:{C1:{brittany} C22:{ship} C34:{judge} C42:{tonne} C52:{ruling} } 5:{C3:{vessel} C5:{tide} C11:{court} C29:{crew} C32:{sinking} C45:{environmentalist} C48:{fishing} C54:{protection} C59:{coastline} C62:{total} }
17	1:{C0:{breton} C6:{oil} C25:{disaster} C29:{spill} } 2:{C3:{business, tourism} C5:{responsibility} C6:{tanker} C7:{precedent} C12:{pollution} C13:{seabird, bird} C19:{ton} C22:{mile} C23:{february, december} } 3:{C6:{fuel} }
18	1:{C0:{bird} } 2:{C9:{oil} C23:{spill} } 3:{C0:{guillemot, gannet, duck, goose} C5:{estuary, inlet, england, bay, sea} } 4:{C0:{puffin} C3:{summer} C6:{hundred, thousand} C8:{authority} C10:{ton, tonne} C15:{wind} C22:{coastline} } 5:{C0:{kingfisher} }
19	1:{C12:{trial} } 2:{C12:{defense} } 3:{C12:{judge, case, prosecutor} } 4:{C10:{oil} C12:{proceeding} } 5:{C6:{france} C12:{court} } 6:{C12:{lawyer} C49:{total} }
20	1:{C0:{sea} C3:{prosecution} C5:{oil, tanker} } 2:{C0:{euro} C3:{case, prosecutor} C5:{ship} } 3:{C0:{france} C3:{trial} C7:{week, year} C10:{company} C13:{plaintiff} C15:{monday} C17:{damage, charge} C18:{pollution} C22:{sea_bird, seabird} C24:{sinking} C25:{beginning} C27:{subsidiary} C29:{check} C35:{disaster} } 4:{C0:{bay} C3:{verdict, court} C5:{fuel} C25:{individual} } 5:{C0:{paris, brittany, coast} }

Tableau 19: Les résultats obtenus par la fonction d'agrégation max($\mu_{tf\text{-}grandsuffisament}$, min($\mu_{s\text{-}grand}$, $\mu_{c\text{-}grand}$))

3.2. Impact des relations sur le calcule de centralité :

Les Figures « Figure 22 et Figure 23 » présentent les résultats obtenus par l'étude sur l'impact des relations sur le calcule de centralité. Ces expérimentations sont réalisée sur deux

collections différentes issues de la collection TREC de NICT « TREC1 » et « TREC7 » Dans ces expérimentations, nous utilisons les termes simples dans l'indexation ainsi que les meilleures méthodes de désambigüisation par centralité pour les documents et les requêtes obtenues précédemment.

Nous utilisons dans les expérimentations notre fonction d'appariement « $c^2*(1+f)*s$ ». Les résultats obtenus pour chaque combinaison de *R* sont présentés sous formes graphiques (Figure 22 pour la collection TREC1, ainsi, la Figure 23 pour la collection TREC7) et sous forme tableau (Tableau 20 et Tableau 21 pour la collection TREC1 et Tableau 22, Tableau 23 pour la collection TREC7). Les valeurs de la précision pour les top-5, top-10 et top-15 documents sont représentées graphiquement dans les figures pour chaque combinaison des relations possibles de R. La combinaison R1(1,1,1,1,1) correspond à l'approche telle que nous l'avons testé précédement c.à.d. nous employons toutes les relations pas seulement en considérant les relations directes mais en plus les relations indirectes entre les concepts.

	R(1,0, 0,0,0)	R(1,0, 0,0,1)	R(1,0, 0,1,0)	R(1,0, 0,1,1)	R(1,0, 1,0,0)	R(1,0, 1,0,1)	R(1,0, 1,1,0)	R(1,0, 1,1,1)
P5	0,4625	0,4792	0,5083	0,4958	0,5292	0,5083	0,5458	0,5208
P10	0,4729	0,4875	0,4896	0,4985	0,4875	0,5042	0,4979	0,5083
P15	0,4597	0,4875	0,4653	0,4931	0,4833	0,4931	0,4764	0,4944
P20	0,4594	0,4771	0,4583	0,4729	0,4635	0,4854	0,4625	0,4833
P30	0,4465	0,4556	0,4458	0,4576	0,4444	0,4646	0,4576	0,4674
P100	0,3633	0,3633	0,3633	0,3633	0,3633	0,3633	0,3633	0,3633
map	0,0959	0,0972	0,0981	0,0993	0,0983	0,0994	0,0998	0,1011

Tableau 20: Impact des différentes relations sur la fonction c^2*f*s pour la collection TREC1

	R(1,1, 0,0,0)	R(1,1, 0,0,1)	R(1,1, 0,1,0)	R(1,1, 0,1,1)	R(1,1, 1,0,0)	R(1,1, 1,0,1)	R(1,1, 1,1,0)	R(1,1, 1,1,1)	R1(1,1, 1,1,1)
P5	0,5083	0,4917	0,5208	0,5125	0,5167	0,4958	0,525	0,5208	0,5375
P10	0,4917	0,5021	0,4958	0,4958	0,4875	0,5083	0,5	0,5125	0,4917
P15	0,4597	0,4819	0,475	0,4861	0,4681	0,4861	0,475	0,4875	0,4861
P20	0,4552	0,4667	0,4677	0,474	0,4667	0,4719	0,4708	0,476	0,4635
P30	0,4451	0,4458	0,4375	0,4514	0,4417	0,4486	0,4403	0,4472	0,4417
P100	0,3633	0,3633	0,3633	0,3633	0,3633	0,3633	0,3633	0,3633	0,3633
map	0,0942	0,0949	0,0967	0,0971	0,0943	0,095	0,0972	0,0975	0,0973

Tableau 21: Impact des différentes relations sur la fonction c^2*f*s pour la collection TREC1

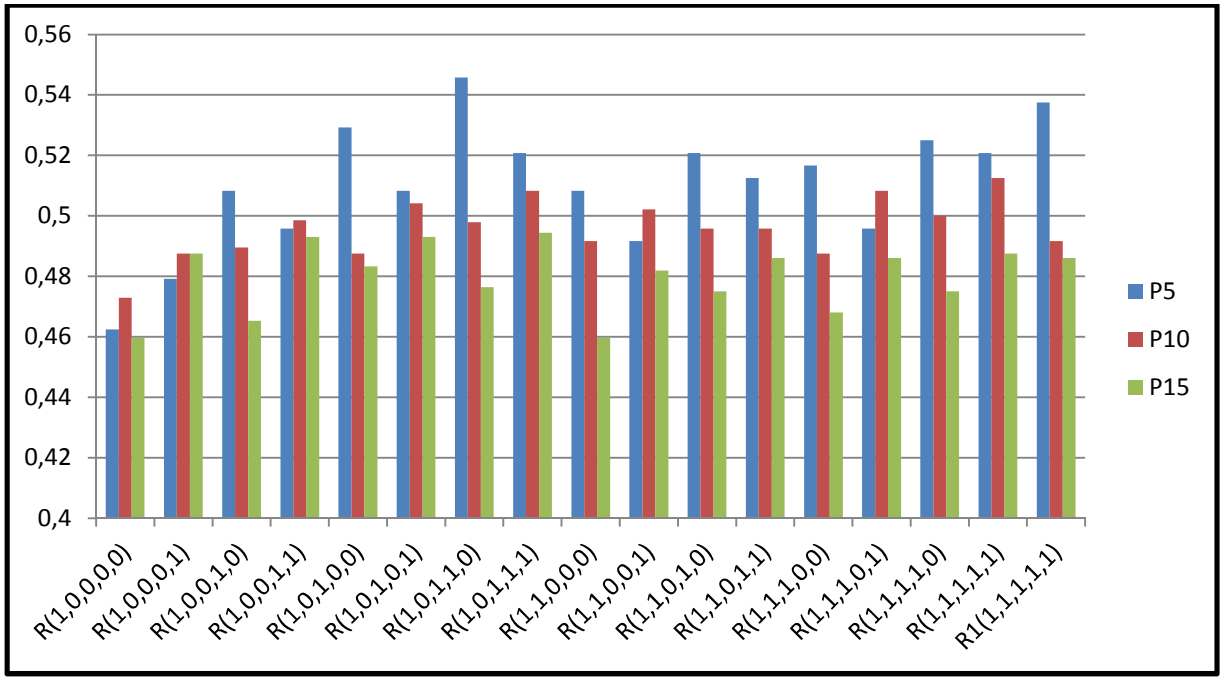


Figure 22: Comparaison graphique de la précision pour les différentes relations sur la collection TREC1

	R(1,0, 0,0,0)	R(1,0, 0,0,1)	R(1,0, 0,1,0)	R(1,0, 0,1,1)	R(1,0, 1,0,0)	R(1,0, 1,0,1)	R(1,0, 1,1,0)	R(1,0, 1,1,1)
P5	0,298	0,3102	0,2612	0,2694	0,3306	0,351	0,302	0,302
P10	0,2878	0,2959	0,2429	0,2571	0,2898	0,302	0,2653	0,2673
P15	0,2571	0,2694	0,234	0,2503	0,2735	0,2884	0,2544	0,2653
P20	0,2408	0,2582	0,2296	0,2439	0,2582	0,2724	0,2408	0,25
P30	0,234	0,2415	0,2272	0,2293	0,2449	0,2531	0,234	0,2408
P100	0,1988	0,1988	0,1988	0,1988	0,1988	0,1988	0,1988	0,1988
map	0,1051	0,1066	0,098	0,0996	0,1093	0,1105	0,1041	0,1034

Tableau 22: Impact des différentes relations sur la fonction c^2*f*s pour la collection TREC7

	R(1,1, 0,0,0)	R(1,1, 0,0,1)	R(1,1, 0,1,0)	R(1,1, 0,1,1)	R(1,1, 1,0,0)	R(1,1, 1,0,1)	R(1,1, 1,1,0)	R(1,1, 1,1,1)	R1(1,1, 1,1,1)
P5	0,2857	0,2939	0,2898	0,298	0,3143	0,3265	0,3102	0,3102	0,2939
P10	0,2878	0,2959	0,2633	0,2857	0,2939	0,3061	0,2796	0,3	0,2796
P15	0,2721	0,2789	0,2463	0,2626	0,2721	0,2857	0,2544	0,2707	0,2667
P20	0,2622	0,2633	0,2418	0,2449	0,2694	0,2724	0,252	0,2571	0,2531
P30	0,2517	0,2524	0,2333	0,2347	0,2578	0,2599	0,2381	0,2367	0,2456
P100	0,1988	0,1988	0,1988	0,1988	0,1988	0,1988	0,1988	0,1988	0,1988
map	0,106	0,1083	0,1014	0,1041	0,1086	0,1106	0,1043	0,106	0,1062

Tableau 23: Impact des différentes relations sur la fonction c^2*f*s pour la collection TREC7

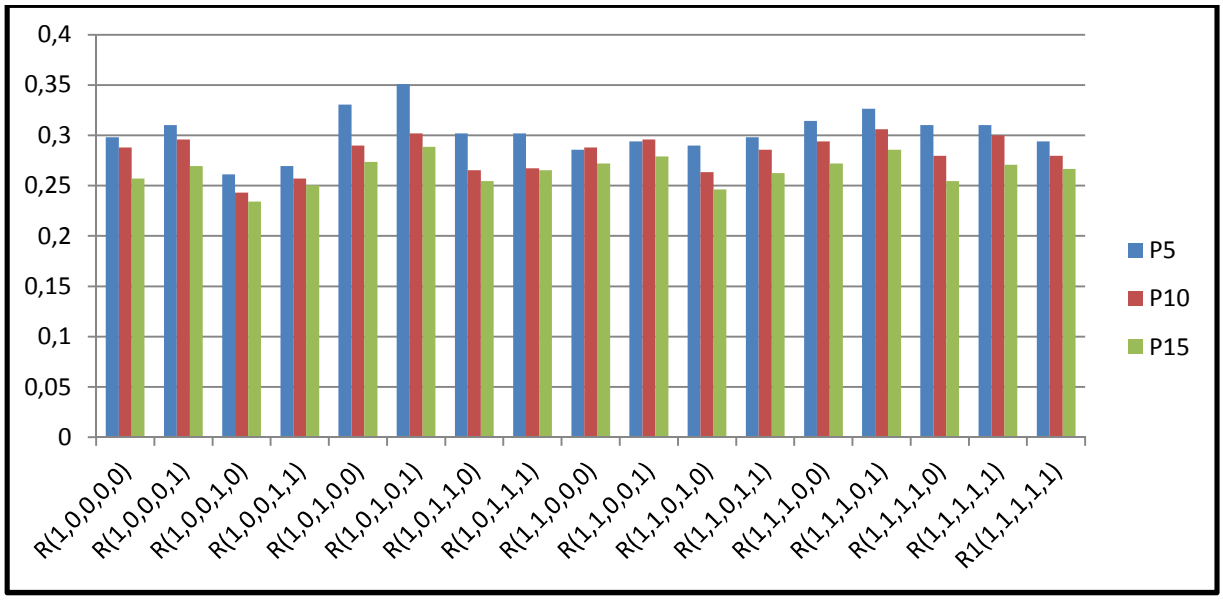


Figure 23: Comparaison graphique de la précision pour les différentes relations sur la collection TREC7

3.3. Impact des relations sur Cf

Nous présentons les résultats obtenus par les deux fonctions de tests sur la collection TREC1. Les résultats de la fonction « Okapi-BM25 basée sur Cf » obtenus pour chaque combinaison de *R* sont présentés sous formes tableaux et graphiques dans Tableau 24, Tableau 25 et la Figure 24 et les résultats de la fonction du modèle de langue basée sur Cf sont présentés sous formes tableaux dans Tableau 26, Tableau 27 et sous forme graphique dans la Figure 25. Chaque colonne de ces tableaux liste la précision de la fonction obtenue en utilisant une certaine combinaison de *R* pour les top-x document (top5, top10, top15…), ensuite les mêmes valeurs de la précision uniquement pour les top-5, top-10 et top-15 documents sont représentés graphiquement dans les figures. La combinaison R1(1,1,1,1,1) indique que nous utilisons toutes les relations, pas seulement en considérons les relations directe mais en plus les relations indirecte entre les concepts.

	R(1,0, 0,0,0)	R(1,0, 0,0,1)	R(1,0, 0,1,0)	R(1,0, 0,1,1)	R(1,0, 1,0,0)	R(1,0, 1,0,1)	R(1,0, 1,1,0)	R(1,0, 1,1,1)
P5	0,4500	0,4333	0,4375	0,4292	0,4458	0,4292	0,4417	0,4292
P10	0,4354	0,4333	0,4438	0,4438	0,4333	0,4313	0,4458	0,4438
P15	0,4222	0,4236	0,4319	0,4319	0,4236	0,4222	0,4319	0,4306
P20	0,4104	0,4135	0,4281	0,4302	0,4135	0,4167	0,4312	0,4333
P30	0,4132	0,4188	0,4215	0,4264	0,4153	0,4222	0,4222	0,4271
P100	0,3633	0,3633	0,3633	0,3633	0,3633	0,3633	0,3633	0,3633
map	0,0928	0,0931	0,0953	0,0956	0,0932	0,0934	0,0957	0,0960

Tableau 24: Impact des différentes relations sur la fonction BM25 pour la collection TREC1

	R(1,1, 0,0,0)	R(1,1, 0,0,1)	R(1,1, 0,1,0)	R(1,1, 0,1,1)	R(1,1, 1,0,0)	R(1,1, 1,0,1)	R(1,1, 1,1,0)	R(1,1, 1,1,1)	R1(1,1, 1,1,1)
P5	0,4875	0,4792	0,4750	0,4708	0,4875	0,4792	0,4750	0,4708	0,4875
P10	0,4729	0,4542	0,4729	0,4646	0,4729	0,4542	0,4708	0,4625	0,4896
P15	0,4667	0,4556	0,4625	0,4528	0,4681	0,4569	0,4625	0,4528	0,4778
P20	0,4604	0,4562	0,4604	0,4521	0,4646	0,4562	0,4625	0,4531	0,4604
P30	0,4278	0,4299	0,4431	0,4458	0,4319	0,4333	0,4465	0,4479	0,4625
P100	0,3633	0,3633	0,3633	0,3633	0,3633	0,3633	0,3633	0.3633	0,3633
map	0,0961	0,0958	0,0982	0,0981	0,0965	0,0962	0,0986	0.0985	0,1043

Tableau 25: Impact des différentes relations sur la fonction BM25 pour la collection TREC1

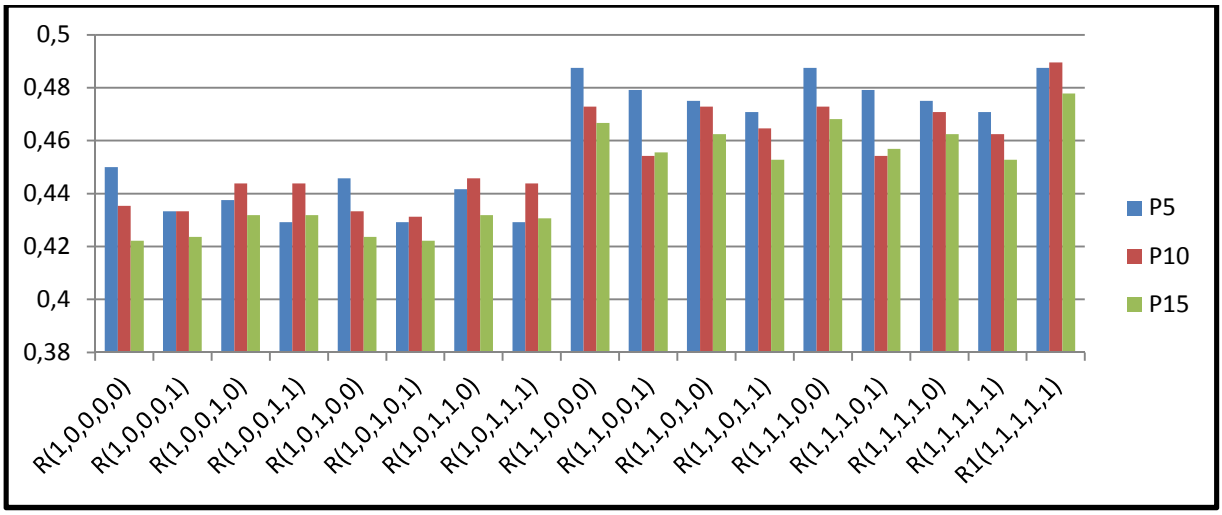


Figure 24: Comparaison graphique de la précision pour les différentes relations sur la collection TREC1 pour la fonction BM25

	R(1,0, 0,0,0)	R(1,0, 0,0,1)	R(1,0, 0,1,0)	R(1,0, 0,1,1)	R(1,0, 1,0,0)	R(1,0, 1,0,1)	R(1,0, 1,1,0)	R(1,0, 1,1,1)
P5	0,4917	0,4958	0,5042	0,5083	0,4917	0,4958	0,5042	0,5083
P10	0,4771	0,4729	0,4813	0,4792	0,4750	0,4708	0,4813	0,4771
P15	0,4542	0,4583	0,4556	0,4583	0,4556	0,4597	0,4597	0,4625
P20	0,4479	0,4490	0,4510	0,4531	0,4510	0,4510	0,4531	0,4542
P30	0,4458	0,4451	0,4493	0,4493	0,4472	0,4465	0,4500	0,4507
P100	0,3633	0,3633	0,3633	0,3633	0,3633	0,3633	0,3633	0,3633
map	0,0980	0,0981	0,0987	0,0988	0,0983	0,0985	0,0990	0,0991

Tableau 26: Impact des différentes relations sur la fonction du Modèle de Langue pour la collection TREC1

	R(1,1, 0,0,0)	R(1,1, 0,0,1)	R(1,1, 0,1,0)	R(1,1, 0,1,1)	R(1,1, 1,0,0)	R(1,1, 1,0,1)	R(1,1, 1,1,0)	R(1,1, 1,1,1)	R1(1,1, 1,1,1)
P5	0,5250	0,5250	0,5083	0,5125	0,5250	0,5250	0,5042	0,5083	0,5542
P10	0,4833	0,4833	0,4792	0,4812	0,4833	0,4833	0,4812	0,4833	0,5104
P15	0,4694	0,4708	0,4750	0,4736	0,4722	0,4736	0,4750	0,4750	0,4958
P20	0,4625	0,4625	0,4604	0,4615	0,4625	0,4625	0,4604	0,4615	0,4917
P30	0,4514	0,4528	0,4528	0,4528	0,4528	0,4542	0,4535	0,4535	0,4625
P100	0,3633	0,3633	0,3633	0,3633	0,3633	0,3633	0,3633	0,3633	0,3633
map	0,0992	0,0992	0,0995	0,0997	0,0993	0,0994	0,0997	0,0998	0,1060

Tableau 27: Impact des différentes relations sur la fonction du Modèle de Langue pour la collection TREC1

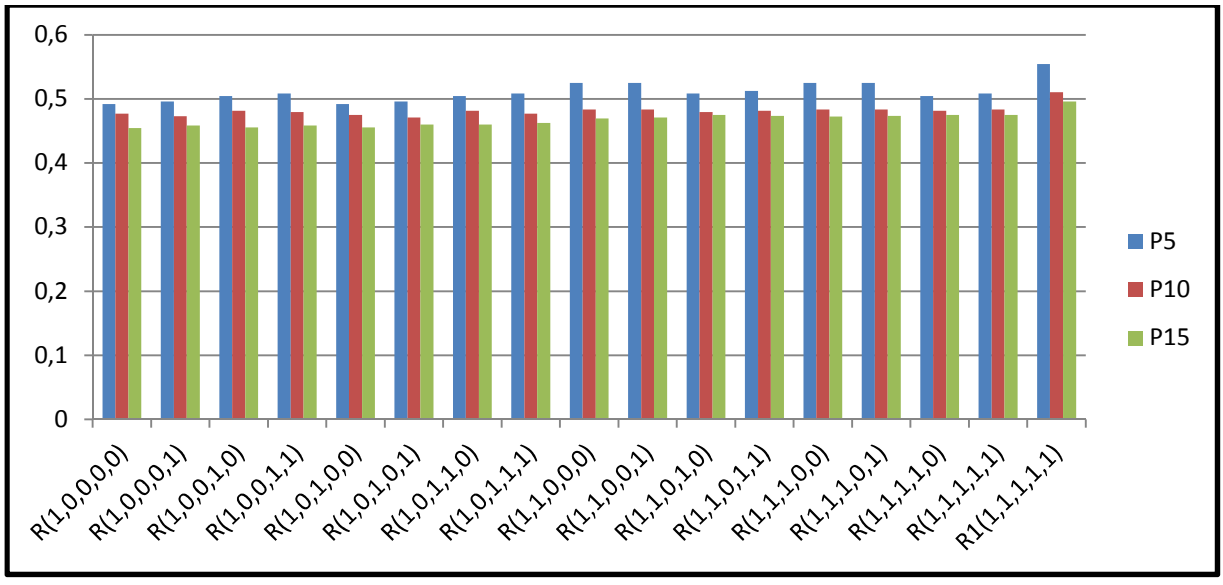


Figure 25: Comparaison graphique de la précision pour les différentes relations sur la collection TREC1 pour le modèle de langue

Nous présentons ensuite les résultats obtenus par les deux fonctions de tests sur la collection TREC7. Les résultats de la fonction « Okapi-BM25 basée sur Cf » obtenus pour chaque combinaison de *R* sont présentés sous formes tableaux et graphiques dans Tableau 28, Tableau 29 et la Figure 26 et les résultats de la fonction du modèle de langue basée sur Cf

sont présentés sous formes tableaux dans Tableau 30, Tableau 31 et sous forme graphique dans la Figure 27. Chaque colonne de ces tableaux liste la précision de la fonction obtenue en utilisant une certaine combinaison de *R* pour les top-x document (top5, top10, top15…), ensuite les mêmes valeurs de la précision uniquement pour les top-5, top-10 et top-15 document sont représentés graphiquement dans les figures. La combinaison R1(1,1,1,1,1) indique que nous utilisons toutes les relations, pas seulement en considérons les relations directe mais en plus les relations indirecte entre les concepts.

	R(1,0, 0,0,0)	R(1,0, 0,0,1)	R(1,0, 0,1,0)	R(1,0, 0,1,1)	R(1,0, 1,0,0)	R(1,0, 1,0,1)	R(1,0, 1,1,0)	R(1,0, 1,1,1)
P5	0,4163	0,4163	0,4204	0,4204	0,4204	0,4204	0,4245	0,4245
P10	0,3755	0,3755	0,3816	0,3816	0,3796	0,3796	0,3816	0,3816
P15	0,3347	0,3333	0,3347	0,3333	0,3388	0,3374	0,3388	0,3374
P20	0,3214	0,3224	0,3224	0,3214	0,3214	0,3214	0,3224	0,3214
P30	0,2844	0,2864	0,2891	0,2905	0,2857	0,2878	0,2905	0,2918
P100	0,1986	0,1986	0,1986	0,1986	0,1986	0,1986	0,1986	0,1986
map	0,1345	0,135	0,7353	0,1352	0,1349	0,1349	0,1357	0,1352

Tableau 28: Impact des différentes relations sur la fonction BM25 pour la collection TREC7

	R(1,1, 0,0,0)	R(1,1, 0,0,1)	R(1,1, 0,1,0)	R(1,1, 0,1,1)	R(1,1, 1,0,0)	R(1,1, 1,0,1)	R(1,1, 1,1,0)	R(1,1, 1,1,1)	R1(1,1, 1,1,1)
P5	0,4449	0,4449	0,4449	0,4449	0,.4367	0,4327	0,4408	0,4367	0,4245
P10	0,3776	0,3796	0,3796	0,3816	0,3755	0,3776	0,3755	0,3776	0,3980
P15	0,3551	0,3551	0,3483	0,3497	0,3497	0,351	0,3442	0,3456	0,3483
P20	0,3224	0,3214	0,3296	0,3276	0,3214	0,3204	0,3296	0,3276	0,3276
P30	0,2932	0,2952	0,2959	0,2966	0,2932	0,2946	0,2973	0,298	0,2939
P100	0,1986	0,1986	0,1986	0,1986	0,1986	0,1986	0,1986	0,1986	0,1986
map	0,1332	0,1341	0,1332	0,1336	0,1333	0,1337	0,1335	0,1336	0,1369

Tableau 29: Impact des différentes relations sur la fonction BM25 pour la collection TREC7

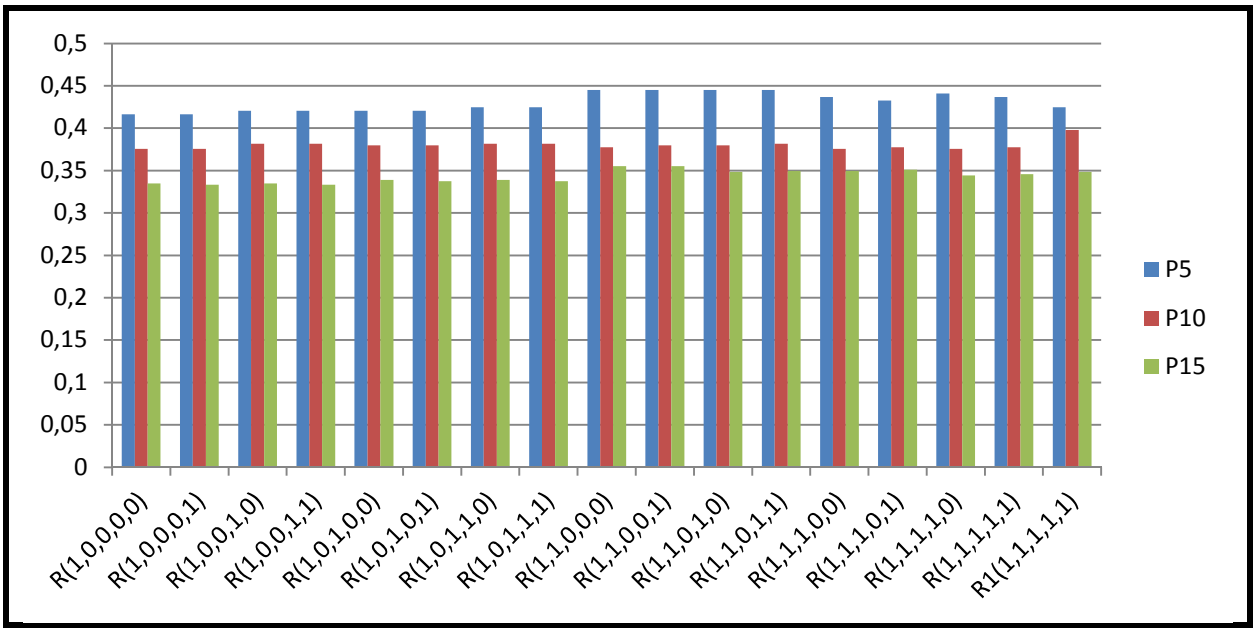


Figure 26: Comparaison graphique de la précision pour les différentes relations sur la collection TREC7 pour la fonction BM25

	R(1,0, 0,0,0)	R(1,0, 0,0,1)	R(1,0, 0,1,0)	R(1,0, 0,1,1)	R(1,0, 1,0,0)	R(1,0, 1,0,1)	R(1,0, 1,1,0)	R(1,0, 1,1,1)
P5	0,4163	0,4204	0,4286	0,4327	0,4163	0,4204	0,4286	0,4327
P10	0,3755	0,3735	0,3694	0,3673	0,3796	0,3776	0,3694	0,3673
P15	0,3401	0,3442	0,3456	0,3469	0,3442	0,3469	0,3497	0,3497
P20	0,3133	0,3173	0,3173	0,3184	0,3163	0,3204	0,3194	0,3204
P30	0,2796	0,2816	0,2796	0,283	0,281	0,283	0,281	0,2844
P100	0,1986	0,1986	0,1986	0,1986	0,1986	0,1986	0,1986	0,1986
map	0,1319	0,1325	0,1328	0,1329	0,1331	0,1336	0,1338	0,1339

Tableau 30: Impact des différentes relations sur la fonction du Modèle de Langue pour la collection TREC7

	R(1,1, 0,0,0)	R(1,1, 0,0,1)	R(1,1, 0,1,0)	R(1,1, 0,1,1)	R(1,1, 1,0,0)	R(1,1, 1,0,1)	R(1,1, 1,1,0)	R(1,1, 1,1,1)	R1(1,1, 1,1,1)
P5	0,4286	0,4286	0,4245	0,4367	0,4367	0,4367	0,4367	0,4449	0,4204
P10	0,3755	0,3755	0,3796	0,3796	0,3755	0,3755	0,3755	0,3755	0,3776
P15	0,3442	0,3469	0,3469	0,3497	0,3442	0,3483	0,351	0,3537	0,3483
P20	0,3173	0,3235	0,3153	0,3184	0,3184	0,3245	0,3163	0,3204	0,3235
P30	0,2905	0,2905	0,2898	0,2912	0,2905	0,2905	0,2918	0,2932	0,2898
P100	0,1986	0,1986	0,1986	0,1986	0,1986	0,1986	0,1986	0,1986	0,1986
map	0,1353	0,1364	0,135	0,1359	0,1359	0,1369	0,1356	0,1362	0,1354

Tableau 31: Impact des différentes relations sur la fonction du Modèle de Langue pour la collection TREC7

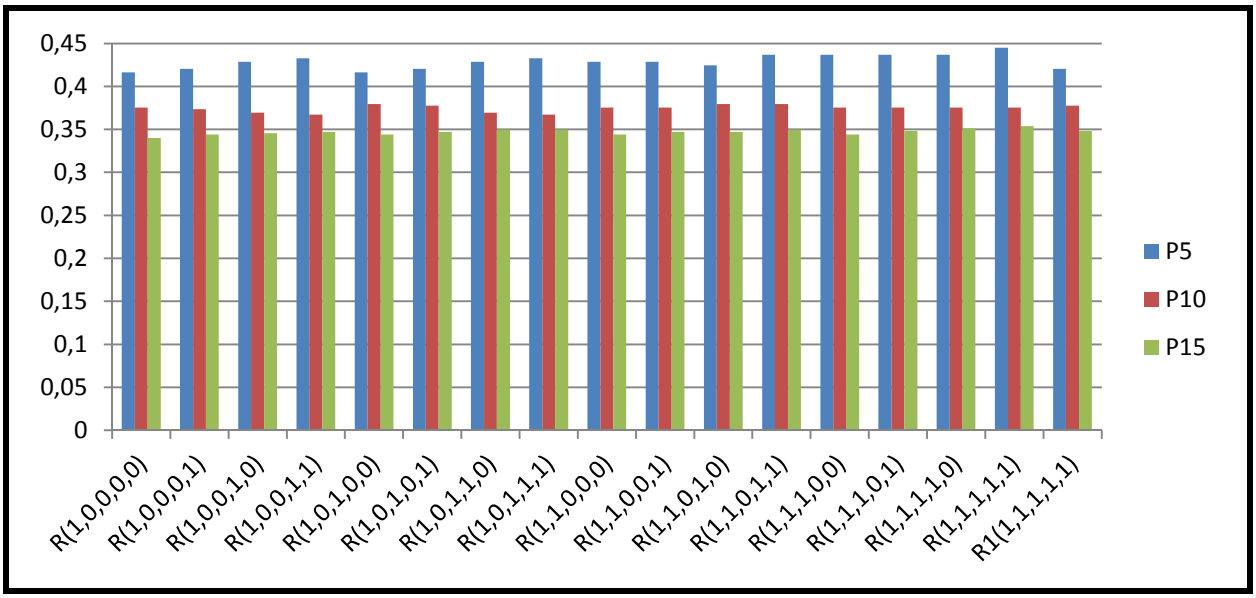


Figure 27: Comparaison graphique de la précision pour les différentes relations sur la collection TREC1 pour le modèle de langue